Descriptors of Superconductivity

Henning Glawe

2018

im
Fachbereich Physik der
Freien Universität Berlin
eingereichte

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium

Supervisor:	Prof. Dr. Eberhard K.U. Groß
Second Referee:	Prof. Dr. Felix von Oppen
Date of defense:	April 29th, 2019

Declaration of Authorship / Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Dissertationsschrift mit dem Titel „Descriptors of Superconductivity" selbständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht.

Desweiteren versichere ich, dass diese Arbeit noch nie in einem anderen Promotionsverfahren vorgelegt wurde.

Datum

Unterschrift (Henning Glawe)

Abstract

The subject of this thesis is the development of theoretical and computational methods to allow for the search of new superconductors via High-Throughput Methods (HTMs)

HTMs test thousands of materials for a desired property. Each test on a material is performed by computational simulation; hence the computational cost of each individual simulation has a strong impact on the global performance of a HTM. While first-principle methods exist and allow for in-depth studies of single superconducting materials, such methods are far too expensive to be applied directly in HTMs.

In order to predict superconductors, one needs to address primarily three challenges, namely the accelerated prediction (i) of new crystalline systems that are thermodynamically or at least dynamically stable and are yet to be synthesized, (ii) of basic electronic properties such as the metallicity and absence of magnetic instabilities, (iii) of the strength of the pairing interaction and hence the possibility of having a high superconducting transition temperature T_c; within this work, we consider conventional electron-phonon superconductivity.

At the core of this work, we focus on challenge (iii) by introducing *descriptors of superconductivity*, that, while accessible at low computational cost, convey sufficient information to select candidate materials for in-depth investigation. Our approach to develop such descriptors is based both on theoretical knowledge and on empirical data, while connections between the former and the latter are demonstrated by models. We implement the *descriptors* numerically on the basis of Kohn-Sham Density Functional Theory (DFT). We perform a high-throughput search, evaluating our *descriptors of superconductivity* on a library of known materials, and identify promising candidate superconductors. For a subset sample of those candidates, we perform expensive full-scale ab initio calculations, validating our *descriptors*.

We address challenge (ii) by developing machine learning (ML) methods with the goal to further accelerate the process by predicting electronic properties directly from the crystal structure of a given material. For this purpose, we introduce a new representation of crystal structures for ML, which takes important symmetries of periodic systems into account. These ML methods are evaluated with the help of the data generated in our high-throughput search.

Finally, we develop a strategy for challenge (i), which allows to predict new crystalline systems via element substitution. By means of statistical analysis performed on a large library of materials, we introduce a new measure for the similarity between chemical elements. This measure by itself can be employed in the prediction of new materials from existing ones. However, we go one step further: based on this data, we propose a new chemical scale, similar in spirit to the well-known Pettifor scale, which provides a one-dimensional order of chemical elements by their chemical similarity.

Contents

Introduction

Since the first observation of superconductivity in Mercury at a critical temperature (T_c) of 4K many superconducting materials have been discovered. However, for a long time, T_c was believed to have an upper bound around 25K [1], up until the discovery of superconductivity at 35K in cuprate perovskites [2]; this limit has been pushed up to 133K, well within the range of liquid Nitrogen cooling, in the same family of materials [3]. However, as these materials are quite brittle ceramics, the production of wires is complex and expensive [4]. Interest on superconductors outside the cuprate class has renewed in 2001 after the discovery that MgB_2, a material known since 1954 [5], becomes superconducting at 39K [6]. This temperature is accessible by cheaper cooling methods than liquid He, while synthesis and mechanical properties of MgB_2 are far less demanding. And again, recently a whole new class, the pnictide superconductors [7], was discovered, which are less brittle than cuprates while exhibiting a maximal T_c of 53K [8]

With each new material with a higher T_c than its predecessors, superconductivity finds more practical applications. Today, superconductors have essential technological applications. The main one is nowadays their use in MRI scanners [9] employed for diagnostics in hospitals. In research, superconducting electromagnets [10] are used wherever strong magnetic fields are required, like in particle accelerators [11], experimental fusion reactors [12], magnetic traps [13–15] and even for energy storage [16]. First superconducting power cables [17] have been made, but the cooling- and manufacturing costs restrict their use to cases where no other solution is viable [18].

Discovering new, better superconductors is therefore an important task, as better superconductors would automatically lower the costs in present applications, while opening up also for new ones. The ultimate goal would be to find a room temperature superconductor, which would have a massive impact on technological development and ultimately, our everyday life.

All families of superconductors known up to now have been discovered by experimental research, with the very recent exception being the discovery of superconductivity at 190 K in H_3S at high pressure (150 GPa), which was first predicted theoretically [19], and only later confirmed experimentally [20]. Many of the experimental discoveries involved chance [21], as in the previously discussed case of MgB_2. However, experimental discovery of new superconductors is an inherently time-consuming task, due to the complex procedure of synthesis. This fact limits the possibility of large scale systematic searches, as done in other fields with experimental high throughput methods (HTM) [22–24].

However HTM can also be applied together with computational methods. During the recent years *ab-initio* high-throughput computational methods have proven to be a powerful and successful tool to predict new materials and to optimize desired material

properties. Phase diagrams of multicomponent crystals [25–27] and alloys [28] have been successfully predicted. High-impact technological applications have been achieved by improving the performance of Lithium based batteries [29–31], by tailoring the non-linear optical response in organic molecules [32] for optical signal processing, by designing desired current-voltage characteristics [33] for photovoltaic materials, by optimizing the electrode transparency and conductivity [34] for solar cell technology, and by screening metals for the highest amalgamation enthalpy [35] to efficiently remove Hg pollutants in coal gasification.

However, HTM can be applied only if the cost of a single material calulation is rather cheap (order of magnitude of 10 hours per material). *Ab-initio* prediction methods for superconductivity [36, 37], although they exist and were able to accurately reproduce the superconducting properties of many systems [38–44], cannot be directly applied in the context of HTM, because they are computationally far too expensive. The reason is that superconductivity is caused by a weak perturbation [45, 46], which by itself is the interplay of competing interactions, which for this reason have to be known with high accuracy. Moreover each of the relevant interactions is computationally expensive: phonon and Coulomb, the established mechanism in conventional superconductors [45–47]; spin fluctuations [48], plasmons [49], polarons [50] and others have been suggested for unconventional superconductors.

The aim of this work is to investigate strategies and solutions for the application of high-throughput methods to the discovery of superconductors.

The main contribution to the field is the formulation of a set of *descriptors of superconductivity*, quantities correlated with superconductivity, but accessible at low computational cost (on the scale of a single Kohn-Sham DFT calculation per material). The descriptors are proposed from a theoretical point of view, focusing on phononic superconductors, since a predictive theory of superconductivity is presently only available for this class. They are evaluated on a large number of existing materials from the inorganic crystal structure database (ICSD) [51, 52]. Based on this data, their predictive power is tested, and a statistical analysis is used to improve the descriptors themselves. This approach leads to the very promising final result, that it is possible to correctly predict superconductivity on a validation set of systems with an accuracy of 80 percent.

A second contribution of this work is the investigation of machine learning methods in order to bypass the computational cost of the Kohn-Sham calculation altogether, which, while being relatively cheap for simple systems, becomes expensive when materials become more complex.

A third contribution is a statistical analysis performed on ICSD, providing data to be applied in structure prediction methods. Based on this statistical data, we establish a scale of chemical similarity that can be used in accelerated materials design. This third contribution, although essential for application of high-throughput methods to superconductivity, goes actually beyond the problem of finding new superconductors, as it can be applied within any field of research on functional materials.

Part I.

Theoretical foundation

Our approach to a computational high-throughput search for superconductors is based on the construction of a set of quantities correlated with superconductivity; we call this set *descriptors of superconductivity* (chapter 5). They are defined by an analysis of first principles theories, and also their numerical evaluation is preformed on this basis. Within this part, these theories of the normal- and superconducting state are reviewed.

Furthermore, in this work we have adopted *machine learning* (ML) techniques to *predict* the descriptors purely from the crystal structure. Such techniques would yield a strong boost to a high-throughput search for superconductors, as predictions typically take fractions of a second. ML techniques are largely unknown within the physics community, therfore a review on the methods relevant to this work is presented in the final chapter of this part.

1. Theory of the normal state

In this chapter, we introduce the theoretical methods we employ for the computational evaluation of normal-state properties of the materials in our high-throughput search. A correct description of the normal state is a key ingredient to any description of the superconducting state; superconductivity usually occurs with a second-order phase transition, therefore the properties of the non-superconducting state uniquely determine the critical transition temperature. All ab-initio calculations in our high-throughput search are performed within the framework of Kohn-Sham density functional theory (KS-DFT), which is outlined in section 1.1. In order to evaluate the superconducting properties for a subset of the materials, calculations of phonon modes, frequencies and the interaction between electrons and phonons are computed within density functional perturbation theory, reviewed in section 1.2.

Within this chapter, we will also set the basis for the construction and evaluation of *descriptors of superconductivity* (chapter 5), a key concept of this work.

1.1. Density Functional Theory

Kohn-Sham density functional theory (KS-DFT) [53] represents one of the most successful computational methods to approach the quantum many-body problem.

The Hohenberg-Kohn theorem [54] (subsection 1.1.1) establishes a one-to-one correspondence between ground state density $n_0(\boldsymbol{r})$ and external potential $v(\boldsymbol{r})$; therefore, any observable, such as the ground-state total energy, can be defined as a functional of $n_0(\boldsymbol{r})$.

Based on this theorem, the Kohn-Sham scheme (section 1.1) introduces an auxiliary non-interacting system having the same ground-state density $n_0(\boldsymbol{r})$ as the fully interacting many-body system; at the same time, the universal *exchange-correlation density functional* $E_{\mathrm{xc}}[n(\boldsymbol{r})]$ is introduced, providing an efficient starting point for the approximation of many-body effects.

Altogether, the Kohn-Sham scheme provides the means to obtain the ground-state density (and by virtue of the Hohenberg-Kohn theorem, in principle any observable) of a quantum many-body system. The computational cost of a numerical solution is within reach of present facilities for many realistic problems, although scaling issues pose a serious problem in an application to systems with more than a few hundred atoms.

1.1.1. Hohenberg-Kohn Theorem

Consider a system of N interacting electrons exposed to an external potential $v(\boldsymbol{r})$ which does not depend on time. All stationary many-body states $|\psi\rangle$ satisfy the time-independent Schrdinger equation

$$\widehat{H}\,|\psi\rangle = E\,|\psi\rangle$$

with the Hamiltonian

$$\widehat{H} = \widehat{T} + \widehat{W} + \widehat{V}.$$
$$\widehat{T} = -\frac{\hbar^2}{2m}\sum_{\alpha}\int \widehat{\Psi}^{\dagger}_{\alpha}(\boldsymbol{r})\nabla^2\widehat{\Psi}_{\alpha}(\boldsymbol{r})\,d^3\boldsymbol{r}$$

is the kinetic energy operator, the electron-electron interaction reads

$$\widehat{W} = \sum_{\alpha,\beta}\iint \widehat{\Psi}^{\dagger}_{\alpha}(\boldsymbol{r})\widehat{\Psi}^{\dagger}_{\beta}(\boldsymbol{r'})w(\boldsymbol{r},\boldsymbol{r'})\widehat{\Psi}_{\beta}(\boldsymbol{r'})\widehat{\Psi}_{\alpha}(\boldsymbol{r})\,d^3\boldsymbol{r}\,d^3\boldsymbol{r'}$$

and the external potential acts via

$$\widehat{V} = \sum_{\alpha}\int \widehat{\Psi}^{\dagger}_{\alpha}(\boldsymbol{r})v(\boldsymbol{r})\widehat{\Psi}_{\alpha}(\boldsymbol{r})\,d^3\boldsymbol{r}.$$

Hohenberg and Kohn [54] prooved the existance of a one-to-one correspondence between external potential $v(\boldsymbol{r})$, ground state wave function $|\psi_0\rangle$ and ground state electron density

$$n_0(\boldsymbol{r}) = \langle\psi_0|\widehat{n}(\boldsymbol{r})|\psi_0\rangle = \left\langle \psi_0\middle|\sum_{\alpha}\widehat{\Psi}^{\dagger}_{\alpha}(\boldsymbol{r})\widehat{\Psi}_{\alpha}(\boldsymbol{r})\middle|\psi_0\right\rangle :$$
$$v(\boldsymbol{r}) \leftrightarrow \psi_0 \leftrightarrow n_0(\boldsymbol{r}).$$

Therefore, the ground-state expectation value of *any* observable $\widehat{O}$ can be formally rewritten as functional of the ground-state density $n_0(\boldsymbol{r})$:

$$\left\langle \psi_0\middle|\widehat{O}\middle|\psi_0\right\rangle = \left\langle \psi_0[n_0(\boldsymbol{r})]\middle|\widehat{O}\middle|\psi_0[n_0(\boldsymbol{r})]\right\rangle = O[n_0(\boldsymbol{r})]. \tag{1.1}$$

The former applies to the ground state total energy E_0 as the ground-state expectation value of the Hamiltonian $\widehat{H}$:

$$\begin{aligned} E_v[n_0(\boldsymbol{r})] &= \left\langle \psi_0[n_0(\boldsymbol{r})]\middle|\widehat{T}+\widehat{W}+\widehat{V}\middle|\psi_0[n_0(\boldsymbol{r})]\right\rangle \\ &= \left\langle \psi_0[n_0(\boldsymbol{r})]\middle|\widehat{T}+\widehat{W}\middle|\psi_0[n_0(\boldsymbol{r})]\right\rangle + \left\langle \psi_0[n_0(\boldsymbol{r})]\middle|\widehat{V}\middle|\psi_0[n_0(\boldsymbol{r})]\right\rangle \\ &:= F_{HK}[n_0(\boldsymbol{r})] + \int n_0(\boldsymbol{r})v(\boldsymbol{r})\,d^3\boldsymbol{r} \end{aligned}$$

with the Hohenberg-Kohn functional $F_{HK}[n_0(\boldsymbol{r})]$. This functional is *universal*, as it does not depend on the particular system characterized by $v(\boldsymbol{r})$ and only depends on the particle-particle interaction $\widehat{W}$.

Given that we knew the an explicit form of $F_{HK}[n(\boldsymbol{r})]$, we could apply the Rayleigh-Ritz variational principle to the total energy functional $E_{v_0}[n(\boldsymbol{r})]$ for a system with a specific external potential v_0 with the corresponding ground-state density $n_0(\boldsymbol{r})$, yielding the Hohenberg-Kohn variational principle:

$$\begin{aligned} &E_0 = E_{v_0}[n_0(\boldsymbol{r})] \text{ and } E_0 < E_{v_0}[n(\boldsymbol{r})]\ \forall n(\boldsymbol{r}) \neq n_0(\boldsymbol{r}) \\ &\implies \delta E_{v_0}[n(\boldsymbol{r})]|_{n(\boldsymbol{r})=n_0(\boldsymbol{r})} = 0. \end{aligned} \tag{1.2}$$

Therefore the ground state density $n_0(\boldsymbol{r})$ and due to (1.1) all observables could be found by a minimization of the total energy with respect to n, without the complexity implied by an explicit minimization with respect to the many-body wave function ψ.

1.1.2. The Kohn-Sham Scheme

The Hohenberg-Kohn variational principle (1.2) suffers from the fact that no explicit form of $F_{HK}[n(\boldsymbol{r})]$ is known. In 1965, Kohn and Sham [53] introduced a scheme that proved to be very useful in practical applications.

It is based on the following assertion: for *each* interacting N-particle system $\widehat{H}$ with potential $v(\boldsymbol{r})$ and ground state density $n_0(\boldsymbol{r})$, there exists a noninteracting system with potential $v_s(\boldsymbol{r})$ with the same ground state density $n_{\mathrm{s},0}(\boldsymbol{r}) \equiv n_0(\boldsymbol{r})$. Its single-particle orbitals φ_n satisfy

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + v_{\mathrm{s}}(\boldsymbol{r})\right)\varphi_n(\boldsymbol{r}) = \epsilon_n \varphi_n(\boldsymbol{r}). \tag{1.3}$$

The ground state of a non-interacting system is the Slater determinant of the N lowest single-particle orbitals ϕ_n with $n = 1 \ldots N$ and $\epsilon_1 < \epsilon_2 < \ldots < \epsilon_N$, therefore its ground state density and energy are

$$n_{\mathrm{s},0}(\boldsymbol{r}) = \sum_{n=1}^{N} |\varphi_n(\boldsymbol{r})|^2 \equiv n_0(\boldsymbol{r}) \qquad E_{\mathrm{s},0} = \sum_{n=1}^{N} \epsilon_n. \tag{1.4}$$

In the case of degenerate single-particle orbitals, the condition $\epsilon_1 < \epsilon_2 < \ldots < \epsilon_N$ obviously cannot be fulfilled. The inclusion of *occupation numbers* γ_n solves this issue and prooves useful in the treatment of metals 1.1.3:

$$n_{\mathrm{s},0}(\boldsymbol{r}) = \sum_{n=1}^{\infty} \gamma_n |\varphi_n(\boldsymbol{r})|^2 \tag{1.5}$$

$$\text{with } \gamma_n = \begin{cases} 1 & \text{for } \epsilon_n < \mu \\ \in [0;1] & \text{for } \epsilon_n = \mu \\ 0 & \text{for } \epsilon_n > \mu \end{cases} \quad \text{and} \quad \sum_{n}^{\infty} \gamma_n = N \tag{1.6}$$

Although the many-body problem has been formally reduced to a single-particle one, the Kohn-Sham scheme is in principle exact: all many-body interactions are still included, but "hidden" inside the potential v_s.

Exchange-correlation Energy

With the ultimate goal of determining v_s in mind, we rewrite the Hohenberg-Kohn energy functional of the interacting many-body system as

$$F_{\mathrm{HK}}[n] := T_{\mathrm{s}}[n] + E_{\mathrm{H}}[n] + E_{\mathrm{xc}}[n]. \tag{1.7}$$

with the kinetic energy of the auxiliary Kohn-Sham orbitals (which are, due to the Hohenberg-Kohn theorem, uniquely defined as a functional of the density)

$$T_{\mathrm{s}}[n] = -\frac{\hbar^2}{2m}\sum_{n=1}^{\infty}\gamma_n\left\langle\varphi_n[n]\middle|\nabla^2\middle|\varphi_n[n]\right\rangle \tag{1.8}$$

and the Hartree energy

$$E_{\mathrm{H}}[n] = \frac{\mathrm{e}^2}{2}\iint\frac{n(\boldsymbol{r})n(\boldsymbol{r'})}{\left|\boldsymbol{r}-\boldsymbol{r'}\right|}\,d^3\boldsymbol{r}d^3\boldsymbol{r'}$$

Rearranging equation (1.7) formally defines the *exchange-correlation functional* as

$$\begin{aligned} E_{\mathrm{xc}}[n] &:= F_{\mathrm{HK}}[n] - T_{\mathrm{s}}[n] - E_{\mathrm{H}}[n] \\ &= (T[n] - T_{\mathrm{s}}[n]) + (W[n] - E_{\mathrm{H}}[n]). \end{aligned} \tag{1.9}$$

As can be seen from this definition, the exchange-correlation energy consists both of the purely non-classical electron-electron interaction contributions (E_{H} corresponds to the classical interaction energy of a charge distribution) and the difference of the interacting and non-interacting kinetic energies. Due to this, E_{xc} provides a reasonable starting point for approximations.

Kohn-Sham potential

The Kohn-Sham potential is determined by applying the variational principle to find the ground state density n_0 of the

$$\begin{aligned} \text{interacting } E[n] &= T_{\mathrm{s}}[n] + V[n] + E_{\mathrm{H}}[n] + E_{\mathrm{xc}}[n] \\ \text{and non-interacting } E_{\mathrm{s}}[n] &= T_{\mathrm{s}}[n] + V_{\mathrm{s}}[n] \end{aligned}$$

systems. Combined with the condition of equal ground state density, $v_{\mathrm{s}}[n(\boldsymbol{r})]$ is determined as

$$v_{\mathrm{s}}[n](\boldsymbol{r}) = v(\boldsymbol{r}) + \mathrm{e}^2\int\frac{n(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}d^3\boldsymbol{r'} + v_{\mathrm{xc}}[n](\boldsymbol{r}) \tag{1.10}$$

with the *exchange-correlation potential*

$$v_{\mathrm{xc}}[n] = \frac{\delta E_{\mathrm{xc}}}{\delta n}. \tag{1.11}$$

By using (1.8) to express $T_{\mathrm{s}}[n]$ in terms of the Kohn-Sham eigenvalues ϵ_n, the total energy of the interacting system can be obtained from

$$E[n] = \sum_{n=1}^{N} \gamma_n \epsilon_n - \frac{\mathrm{e}^2}{2} \iint \frac{n(\boldsymbol{r}) n(\boldsymbol{r'})}{|\boldsymbol{r} - \boldsymbol{r'}|}\, d^3\boldsymbol{r}\, d^3\boldsymbol{r'} + E_{\mathrm{xc}}[n] - \int v_{\mathrm{xc}}(\boldsymbol{r}) n(\boldsymbol{r})\, d^3\boldsymbol{r}. \tag{1.12}$$

Local density approximation (LDA)

An ingredient still missing up to this point is an approximation of $E_{\mathrm{xc}}[n]$ and $v_{\mathrm{xc}}[n]$. In the original paper [53], Kohn and Sham suggested the *local density approximation (LDA)*

$$\begin{aligned} E_{\mathrm{xc}}^{\mathrm{LDA}}[n] &:= \int \epsilon_{\mathrm{xc}}^{\mathrm{HEG}}(n(\boldsymbol{r})) n(\boldsymbol{r}) d^3\boldsymbol{r} \\ v_{\mathrm{xc}}^{\mathrm{LDA}}[n](\boldsymbol{r}) &= \epsilon_{\mathrm{xc}}^{\mathrm{HEG}}(n(\boldsymbol{r})) + n \frac{d}{d\,n} \epsilon_{\mathrm{xc}}^{\mathrm{HEG}}(n) \Big|_{n=n(\boldsymbol{r})}, \end{aligned} \tag{1.13}$$

where the function $\epsilon_{\mathrm{xc}}^{\mathrm{HEG}}(n)$ is determined from the exchange-correlation energy of the homogenous electron gas with the (uniform) density n. Exchange-correlation-energies at various densities can be determined with high precision from Quantum-Monte-Carlo (QMC) calculations. LDA functionals, such as the Perdew-Zunger functional [55], provide parameterizations interpolating $\epsilon_{\mathrm{xc}}^{\mathrm{HEG}}(n)$ from finite sets of QMC results.

In the limit of homogenous systems, LDA is exact; although originally only considered useful for the case of slowly varying densities, calculations emloying LDA yield reasonable results for a very wide range of systems.

Outline of the solution of the Kohn-Sham equations

The Kohn-Sham equation (1.3, 1.5, 1.6) is a nonlinear Schrdinger equation, where the potential $v_s[n]$ depends on the eigengenstates φ_n via the density n. A self-consistent approach is taken to solve them:

1. Initialize the set $\{\varphi_n\}^{(0)}$ as an initial guess
2. Compute $n_s(\boldsymbol{r})$ from $\{\varphi_n\}^{(i-1)}$ via (1.5, 1.6)
3. Solve (1.3) to obtain $\{\varphi_n\}^{(i)}$
4. Compute $E_{\mathrm{s}}^{(i)} = E_{\mathrm{s}}[n^{(i)}]$
5. Continue at 2 until reaching self-consistency, where $E_{\mathrm{s}}^{(i)} \equiv E_{\mathrm{s}}^{(i-1)}$

1.1.3. Kohn-Sham scheme applied to periodic solids

Description of periodic solids

Although crystal lattices and their formal description are a well-established topic, notation in literature varies. In this subsection, a brief summary of the quantities describing crystal structures relevant in the context of this thesis is presented, with the main intent of fixing the notation.

Bravais vectors and basis In solid-state theory, a crystal is described as an *infinite* set $\mathcal{C}$ of atoms following a regular distribution pattern:

$$\begin{aligned} \mathcal{C} &= \{(\boldsymbol{\tau}_I + \boldsymbol{R}, Z_I)\} & I &= 1 \dots N,\ \boldsymbol{R} \in \mathcal{R} \\ \mathcal{R} &= \{(r_1\boldsymbol{a}_1 + r_2\boldsymbol{a}_2 + r_3\boldsymbol{a}_3)\} & \forall (r_1, r_2, r_3) &\in \mathbb{Z}. \end{aligned}$$

The set of atoms $\{(\boldsymbol{\tau}_i, Z_i)\} : i = 1 \dots N$ is called the *basis* of the crystal. Vectors $\boldsymbol{a}_{1,2,3}$ are called *Bravais vectors* and describe the periodicity: the system is invariant under translation by any of their integer linear combinations $\boldsymbol{R}$, which are called *vectors of the direct lattice.*

Wigner-Seitz cell A *primitive unit cell* is defined as a volume containing one lattice point, with the constraint that the whole space can be gaplessly filled when applying all translations $R \in \mathcal{R}$. The *Wigner-Seitz cell* is the primitive unit cell described by

$$\boldsymbol{c} = i\boldsymbol{a_1} + j\boldsymbol{a_2} + k\boldsymbol{a_3} \quad \forall (i,j,k) \in \left]-\tfrac{1}{2}; \tfrac{1}{2}\right].$$

Conventionally, the coordinates of the nuclei $\boldsymbol{\tau}_I$ are chosen to be compatible to the Wigner-Seitz condition and sharing the point symmetry of the direct lattice.

Reciprocal lattice and 1st Brillouin Zone Corresponding to the direct lattice $\boldsymbol{a}_{1,2,3}$, the *reciprocal lattice* is defined as

$$\boldsymbol{G} \in \mathcal{G},\ \mathcal{G} = \{(i\boldsymbol{b}_1 + j\boldsymbol{b}_2 + k\boldsymbol{b}_3)\}\ \forall i,j,k \in \mathbb{Z}$$

where

$$\boldsymbol{b}_1 = \frac{2\pi \boldsymbol{a}_2 \times \boldsymbol{a}_3}{\boldsymbol{a}_1 \cdot (\boldsymbol{a}_2 \times \boldsymbol{a}_3)},\ \boldsymbol{b}_2 = \frac{2\pi \boldsymbol{a}_3 \times \boldsymbol{a}_1}{\boldsymbol{a}_1 \cdot (\boldsymbol{a}_2 \times \boldsymbol{a}_3)},\ \boldsymbol{b}_3 = \frac{2\pi \boldsymbol{a}_1 \times \boldsymbol{a}_2}{\boldsymbol{a}_1 \cdot (\boldsymbol{a}_2 \times \boldsymbol{a}_3)}$$
$$\Rightarrow\ \boldsymbol{b}_i \cdot \boldsymbol{a}_j = 2\pi\delta_{ij}.$$

Direct and reciprocal lattice are related by a Fourier transformation, vectors $\boldsymbol{G}$ are wave vectors of plane waves with the periodicity of the direct lattice:

$$\mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot(\boldsymbol{r}+\boldsymbol{R})} = \mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{r}}.$$

As in the case of the direct lattice, a Wigner-Seitz primitive cell can be constructed for the reciprocal lattice, which is called the *1st Brillouin zone (BZ)*:

$$\{i\boldsymbol{b_1} + j\boldsymbol{b_2} + k\boldsymbol{b_3} \quad \forall (i,j,k) \in \left]-\tfrac{1}{2}; \tfrac{1}{2}\right]\}.$$

Bloch Theorem For the time being, we include the effect of the nuclei by treating their potential as the external potential $v(\boldsymbol{r})$ in the single-particle electronic Hamiltonian:

$$\widehat{h} = -\frac{\hbar^2}{2m}\nabla^2 + v(\boldsymbol{r}) \quad \text{with } v(\boldsymbol{r}+\boldsymbol{R}) = v(\boldsymbol{r})\,\forall \boldsymbol{R} \in \mathcal{R}$$

due to the periodicity of the crystal structure. The Bloch theorem states that eigenstates of this single-particle Hamiltonian have the form

$$\varphi_{n\boldsymbol{k}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} u_{n\boldsymbol{k}}(\boldsymbol{r}) \quad \text{with } u_{n\boldsymbol{k}}(\boldsymbol{r}+\boldsymbol{R}) = u_{n\boldsymbol{k}}(\boldsymbol{r})\,\forall \boldsymbol{R} \in \mathcal{R}. \tag{1.14}$$

The vector $\hbar\boldsymbol{k}$ is called the *crystal momentum*; it does not correspond to the real momentum of an electron, and $\boldsymbol{k}$ solely describes the translational properties of the wave function. Index n is called the *band index* and enumerates states with identical $\boldsymbol{k}$.

Band Structure

Bloch's theorem can be applied to the Kohn-Sham system: the effective single-particle potential

$$v_\mathrm{s}[n](\boldsymbol{r}) = v(\boldsymbol{r}) + v_\mathrm{H}[n](\boldsymbol{r}) + v_\mathrm{xc}[n](\boldsymbol{r})$$

shares the translational invariance

$$v_\mathrm{s}[n](\boldsymbol{r}+\boldsymbol{R}) = v_\mathrm{s}[n](\boldsymbol{r})$$

of the external potential: the phase $\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}$ cancels in the evaluation of the density(1.5), therefore both Hartree and exchange-correlation potentials are invariant under translations by direct lattice vectors $\boldsymbol{R}$.

The Kohn-Sham equation for Bloch states under a periodic v_s reads

$$\begin{aligned}
\left(-\frac{\hbar^2}{2m}\nabla^2 + v_\mathrm{s}(\boldsymbol{r})\right)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} u_{n\,\boldsymbol{k}}(\boldsymbol{r}) &= \epsilon_{n\,\boldsymbol{k}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} u_{n\,\boldsymbol{k}}(\boldsymbol{r}) \\
\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\left(\frac{\hbar^2}{2m}\left(\frac{1}{\mathrm{i}}\nabla + \boldsymbol{k}\right)^2 + v_\mathrm{s}(\boldsymbol{r})\right) u_{n\,\boldsymbol{k}}(\boldsymbol{r}) &= \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\epsilon_{n\,\boldsymbol{k}} u_{n\,\boldsymbol{k}}(\boldsymbol{r}) \\
\Longrightarrow \left(\frac{\hbar^2}{2m}\left(\frac{1}{\mathrm{i}}\nabla + \boldsymbol{k}\right)^2 + v_\mathrm{s}(\boldsymbol{r})\right) u_{n\,\boldsymbol{k}}(\boldsymbol{r}) &= \epsilon_{n\,\boldsymbol{k}} u_{n\,\boldsymbol{k}}(\boldsymbol{r}).
\end{aligned} \tag{1.15}$$

As $\boldsymbol{k}$ is reduced to a parameter in the Hamiltonian, the corresponding eigenvalues $\epsilon_{n\,\boldsymbol{k}}$ become a continuous function $\epsilon_n(\boldsymbol{k})$ in the limit of an infinite number of lattice points. The set $\{\epsilon_n(\boldsymbol{k})\,\forall n, \boldsymbol{k}\}$ is referred to as the *Kohn-Sham band structure* of the crystal. Taking into account both $v_\mathrm{s}(\boldsymbol{r}+\boldsymbol{R}) = v_\mathrm{s}(\boldsymbol{r})$ and $u_{n\,\boldsymbol{k}}(\boldsymbol{r}+\boldsymbol{R}) = u_{n\,\boldsymbol{k}}(\boldsymbol{r})$, (1.15) can be treated as an eigenvalue problem within a single unit cell, solved independently for each $\boldsymbol{k}$, with the band index n enumerating the solutions. Due to the fact that the *reciprocal*

lattice is invariant under translations by $\boldsymbol{G}$, the sets of wave functions and eigenvalues for points $\boldsymbol{k}$ and $\boldsymbol{k}+\boldsymbol{G}$ must be identical:

$$\left\{\varphi_{n\,(\boldsymbol{k}+\boldsymbol{G})}(\boldsymbol{r})\right\} = \left\{\varphi_{n\,(\boldsymbol{k})}(\boldsymbol{r})\right\} \qquad \left\{\epsilon_{n\,(\boldsymbol{k}+\boldsymbol{G})}\right\} = \left\{\epsilon_{n\,(\boldsymbol{k})}\right\}.$$

Therefore, the solution of the Kohn-Sham equations can be restricted to the 1BZ without any loss of information.

Discretization in reciprocal space is employed in practical calculations: the Kohn-Sham-Hamiltonian is diagonalized only for a finite set of $\boldsymbol{k}$ points. In the context of this thesis, these points lie in a regular grid of points in the reciprocal unit cell [56], and the *smearing technique* (Sec. 1.1.3) is used for interpolation when performing BZ integrals.

Density of states and Fermi energy

Some quantities, such as a subset of those introduced in the later chapters, depend only on an overall picture of the band structure. The *density of states* (DOS) is introduced as an integral form of the band structure, which just depends on the energy ϵ:

$$g(\epsilon) = 2\sum_n \frac{1}{(2\pi)^3}\int \delta(\epsilon - \epsilon_n(\boldsymbol{k}))\, d^3\boldsymbol{k}. \tag{1.16}$$

With the help of the DOS, one can define the *Fermi level* ϵ_F as:

$$N = \int_{-\infty}^{\epsilon_\mathrm{F}} g(\epsilon) d\epsilon,$$

where N is the number of electrons in the unit volume. In other words: the ground state of N Bloch-electrons is constructed by occupying all states $\varphi_{n\,\boldsymbol{k}}$ for each $\boldsymbol{k}$, where the corresponding eigenvalue $\epsilon_{n\,\boldsymbol{k}}$ lies below the Fermi energy ϵ_F.

Depending on the band structure $\{\epsilon_n(\boldsymbol{k})\}$ relative to ϵ_F, materials are divided into two classes with fundamentally different physical properties: Insulators and metals.

Insulators

In insulators, there exists a clean separation of the band structure into *valence bands* ($\epsilon_n^{(\mathrm{v})}(\boldsymbol{k}) < \epsilon_\mathrm{F}\,\forall\boldsymbol{k}$) and *conduction bands* ($\epsilon_n^{(\mathrm{c})}(\boldsymbol{k}) > \epsilon_\mathrm{F}\,\forall\boldsymbol{k}$), implying that any band $\epsilon_n(\boldsymbol{k})$ is either completely filled or empty. The size of the band gap

$$\epsilon_\mathrm{G} = \min_{n\,\boldsymbol{k}} \epsilon_n^{(\mathrm{c})}(\boldsymbol{k}) - \max \epsilon_n^{(\mathrm{v})}(\boldsymbol{k})$$

gives rise to the destinction between semiconductors ($\epsilon_\mathrm{G} \simeq k_\mathrm{B}T_\mathrm{room}$) and insulators ($\epsilon_\mathrm{G} > k_\mathrm{B}T_\mathrm{room}$). Strictly speaking, ϵ_F is not uniquely defined in both cases, but may lie anywhere within the gap.

Metals

In a metal *partially filled* bands m with $\epsilon_m(\boldsymbol{k}_\mathrm{F}) = \epsilon_\mathrm{F}$ exist. The surface in reciprocal space decribed by the *Fermi wave vectors* $\{\boldsymbol{k}_\mathrm{F}\}$ is called the *Fermi surface*. Electronic transport properties such as the electrical and thermal conductivity of metals are mainly governed by the states within the Fermi surface, and one needs an accurate description of particularly these states. Due to this fact, the convergence of the total density with respect to the number of $\boldsymbol{k}$ points needed in every step of the self-consistent solution of the Kohn-Sham equations is very slow.

The smearing technique provides a compromise between accuracy and computational effort: each Kohn-Sham energy level is broadened by a smearing function

$$S_\sigma(\epsilon) = \frac{1}{\sigma} S\left(\frac{\epsilon}{\sigma}\right) \qquad \text{where } \int S(x)\,\mathrm{d}x \equiv 1$$

of width σ, entering the DOS in place of the delta function:

$$g(\epsilon) = 2\sum_n \frac{1}{(2\pi)^3} \int S_\sigma(\epsilon - \epsilon_n(\boldsymbol{k}))\, d^3\boldsymbol{k}.$$

σ denotes the smearing width. The occupation numbers γ_n originally developed in the context of degenerate Kohn-Sham states (1.5) are reconsidered and computed from the smoothed step function $\theta_S(\epsilon) = \int_{-\infty}^{\epsilon} S(\epsilon')d\epsilon'$ corresponding to $S_\sigma(\epsilon)$:

$$n_\mathrm{s}(\boldsymbol{r}) = \sum_{n=1}^{\infty} \int \theta_S\left(\frac{\epsilon_\mathrm{F} - \epsilon_n(\boldsymbol{k})}{\sigma}\right) |u_{n\,\boldsymbol{k}}(\boldsymbol{r})|^2\, d^3\boldsymbol{k}. \tag{1.17}$$

Effectively, the smearing technique can be interpreted as a method of sampling the Fermi surface employing interpolation based on the states energetically close to it. Therefore, convergence in metals is achieved with significantly coarser $\boldsymbol{k}$-point samplings when the smearing technique is employed.

1.1.4. Summary

Kohn-Sham density functional theory is a computationally efficient and practically successful scheme for the approximate solution of the quantum many-body problem, both applicable within finite systems, such as atoms and molecules, and periodic crystals.

In the next section, an extension giving access to phonons and their coupling to electrons is reviewed.

1.2. Phonons from Density Functional Perturbation Theory

In this section, a perturbative approach applied to Kohn-Sham DFT , termed *Density functional perturbation theory* (DFPT) is reviewed, which enormously faciliates computational access to lattice vibrational properties.

This section starts from a short introduction of the Born-Oppenheimer approximation and phonons, then gradually introduces DFPT and its application to lattice vibrational properties.

The closing subsection 1.2.4 reviews a perturbative approach for evaluating the interaction between electrons and phonons, which is the key ingredient to any theoretical description of superconductivity.

1.2.1. Born-Oppenheimer Approximation

Up to this point, the nuclei have entered just via the external potential $v(\boldsymbol{r})$, treating them as pinned at fixed positions and without any intrinsic degrees of freedom. However, many phenomena, the most prominent being superconductivity, covered in the context of this thesis, are governed by nuclear motion and its effect on the electronic states. The starting point of this discussion is the full many-body electron-nuclear Hamiltonian

$$\widehat{H} = \widehat{T}^{\mathrm{e}} + \widehat{U}^{\mathrm{ee}} + \widehat{T}^{\mathrm{n}} + \widehat{U}^{\mathrm{nn}} + \widehat{U}^{\mathrm{en}}.$$

$\widehat{T}^{\mathrm{e}}$ and $\widehat{T}^{\mathrm{n}}$ are the kinetic energy operators of the electrons and nuclei, $\widehat{U}^{\mathrm{ee}}$ and $\widehat{U}^{\mathrm{nn}}$ denote the electron-electron and nucleus-nucleus interaction, while $\widehat{U}^{\mathrm{en}}$ describes the electron-nuclear interaction, i.e. the part we treated as the external potential $v(\boldsymbol{r})$ before. Based on the ratio of masses between electrons and nuclei, the Born-Oppenheimer approximation assumes an adiabatic behaviour of electrons, i.e. for each set of fixed nuclear coordinates $\underline{\boldsymbol{R}} = \{\boldsymbol{R_I}\}$, electrons assume an eigenstate $\boldsymbol{\Psi}_{\underline{\boldsymbol{R}},n}$ of the electronic Born-Oppenheimer Hamiltonian

$$\widehat{H}_{\mathrm{BO}} = \widehat{T}^{\mathrm{e}} + \widehat{U}^{\mathrm{ee}} + E^{\mathrm{nn}}(\underline{\boldsymbol{R}}) + \widehat{U}^{\mathrm{en}}(\underline{\boldsymbol{R}}),$$

with the bare nuclear electrostatic energy

$$E^{\mathrm{nn}}(\underline{\boldsymbol{R}}) = \frac{\mathrm{e}^2}{2} \sum_{I \neq J} \frac{Z_I Z_J}{|R_I - R_J|}$$

and the electron-ion interaction

$$\widehat{U}^{\mathrm{en}}(\underline{\boldsymbol{R}}) = \sum_I \int \widehat{\Psi}^\dagger(\boldsymbol{r}) \frac{-Z_I \mathrm{e}^2}{|\boldsymbol{r} - \boldsymbol{R}_I|} \widehat{\Psi}(\boldsymbol{r})\, d^3\boldsymbol{r} = \int \widehat{n}(\boldsymbol{r}) v^{\mathrm{en}}_{\underline{\boldsymbol{R}}}(\boldsymbol{r})\, d^3\boldsymbol{r}$$

with the nuclear charges Z_I. The electronic Schrödinger equation then reads

$$\widehat{H}_{\mathrm{BO}}(\underline{\boldsymbol{R}}) \Psi_{n\,\underline{\boldsymbol{R}}}(\boldsymbol{r}) = E_n(\underline{\boldsymbol{R}}) \Psi_{n\,\underline{\boldsymbol{R}}}(\boldsymbol{r})$$

The total energies $E_n(\underline{\boldsymbol{R}})$ comprise *Born-Oppenheimer potential energy surfaces*, one for each electronic many-body state (enumerated by index n); the surface originating from the electronic ground states $n = 0$ is called the *ground-state Born-Oppenheimer energy surface* $E_0(\underline{\boldsymbol{R}})$, which enters a purely nuclear Schrödinger-like equation:

$$\left(\widehat{T}^{\mathrm{n}} + E_0(\underline{\boldsymbol{R}})\right) \Phi(\underline{\boldsymbol{R}}) = \varepsilon \Phi(\underline{\boldsymbol{R}}). \tag{1.18}$$

Therefore the Born-Oppenheimer approximation allows to split the electronic and nuclear problems. It assumes, that all nuclear interations are instantanous, and due to the adiabaticity, that no transistions of the electronic state are induced by the nuclear motion. The resulting electron-nuclear wave function is then simply a product of the electronic and nuclear states $\Phi(\underline{\boldsymbol{R}}) \otimes \Psi_{\underline{\boldsymbol{R}}}(\underline{\boldsymbol{r}})$.

Interatomic force constants

The force acting on nucleus I in the geometry $\underline{\boldsymbol{R}}$ is given by the first derivative of $E_0(\underline{\boldsymbol{R}})$ with respect to $\boldsymbol{R}_I$. Due to the Hellman-Feynman-Theorem, it can be expressed as

$$\boldsymbol{F}_I = -\frac{\partial E_0(\underline{\boldsymbol{R}})}{\partial \boldsymbol{R}_I} = -\left\langle \Psi_{0\,\underline{\boldsymbol{R}}} \middle| \frac{\partial \widehat{H}_{\mathrm{BO}}(\underline{\boldsymbol{R}})}{\partial \boldsymbol{R}_I} \middle| \Psi_{0\,\underline{\boldsymbol{R}}} \right\rangle$$

The *equilibrium positions* of the nuclei are defined as the positions $\underline{\boldsymbol{R}}$ with vanishing forces acting on the nuclei, i.e. $\boldsymbol{F}_I = 0 \,\forall I$. Assuming small motion of the nuclei around these positions, the *harmonic approximation* can be applied to determine the phonon frequencies by diagonalizing the Hessian matrix of $E(\underline{\boldsymbol{R}})$:

$$\det \left| \frac{1}{\sqrt{M_I M_J}} \frac{\partial^2 E(\underline{\boldsymbol{R}})}{\partial \boldsymbol{R}_I \partial \boldsymbol{R}_J} - \omega^2 \right| = 0. \tag{1.19}$$

The elements of this Hessian matrix are called *inter-atomic force constants* (IFC), and expand as:

$$\begin{aligned} \frac{\partial^2 E(\underline{\boldsymbol{R}})}{\partial \boldsymbol{R}_I \partial \boldsymbol{R}_J} &= \frac{\partial \boldsymbol{F}_I(\underline{\boldsymbol{R}})}{\partial \boldsymbol{R}_J} \\ &= \int \frac{\partial n_{0\,\underline{\boldsymbol{R}}}(\boldsymbol{r})}{\partial \boldsymbol{R}_J} \frac{\partial v^{\mathrm{en}}_{\underline{\boldsymbol{R}}}(\boldsymbol{r})}{\partial \boldsymbol{R}_I} \mathrm{d}^3\boldsymbol{r} + \int n_{0\,\underline{\boldsymbol{R}}}(\boldsymbol{r}) \frac{\partial^2 v^{\mathrm{en}}_{\underline{\boldsymbol{R}}}(\boldsymbol{r})}{\partial \boldsymbol{R}_I \partial \boldsymbol{R}_J} \mathrm{d}^3\boldsymbol{r} + \frac{\partial^2 E^{\mathrm{nn}}(\underline{\boldsymbol{R}})}{\partial \boldsymbol{R}_I \partial \boldsymbol{R}_J}, \end{aligned}$$

which implies that the IFC can be determined knowing only the ground state electronic density for a given nuclear configuration $n_{0\,\underline{\boldsymbol{R}}}(\boldsymbol{r})$ and its linear response $\frac{\partial n_{0\,\underline{\boldsymbol{R}}}(\boldsymbol{r})}{\partial \boldsymbol{R}_J}$ to a change in the nuclear positions.

Dynamical matrix

In the infinite solid, the index I of an atom in the definition of the interatomic force constants (1.19) can be interpreted, with the help of periodic boundary conditions, as a collective index $I = (l, s)$, where s enumerates the members of the primitive cell, while l indentifies the *periodic image* of the cell generated via the periodic boundary conditions. With the introduction of this notation, the atomic coordinate $\boldsymbol{R}_I$ can be decomposed as

$$\boldsymbol{R}_I = \tilde{\boldsymbol{R}}_l + \boldsymbol{\tau}_s + \boldsymbol{u}_s(l),$$

where $\tilde{\boldsymbol{R}}_l \in \mathcal{R}$ denotes the origin coordinates of cell image l, $\boldsymbol{\tau}_s$ are the coordinates of atom s within the primitive cell, and $\boldsymbol{u}_s(l)$ covers the any deviation.

Given two atoms $I = (l, s)$ and $J = (m, t)$, displaced by $u_s^\alpha(l)$ and $u_t^\beta(m)$, where α and β denote the cartesian components, the interatomic force constants read

$$C_{st}^{\alpha\beta}(l,m) := \frac{\partial^2}{\partial u_s^\alpha(l)\partial u_t^\beta(m)} = C_{st}^{\alpha\beta}(\tilde{\boldsymbol{R}}_l, \tilde{\boldsymbol{R}}_m),$$

where, due to the translational invariance of the IFC, the dependence on the image origins $\tilde{\boldsymbol{R}}_l$ and $\tilde{\boldsymbol{R}}_m$ is solely through their difference

$$= C_{st}^{\alpha\beta}(\tilde{\boldsymbol{R}}_l - \tilde{\boldsymbol{R}}_m) =: C_{st}^{\alpha\beta}(\tilde{\boldsymbol{R}}).$$

Performing a Fourier transformation in $\tilde{\boldsymbol{R}}$

$$\tilde{C}_{st}^{\alpha\beta}(\boldsymbol{q}) = \frac{\partial^2 E_0}{\partial u_s^{\alpha*}(\boldsymbol{q})\partial u_t^\beta(\boldsymbol{q})},$$

where $\boldsymbol{u}_s(\boldsymbol{q})$ is defined via the distortion pattern

$$\boldsymbol{R}_I\left[\boldsymbol{u}_s(\boldsymbol{q})\right] := \tilde{\boldsymbol{R}}_{\boldsymbol{l}} + \boldsymbol{\tau}_s + \boldsymbol{u}_s(\boldsymbol{q})\mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\tilde{\boldsymbol{R}}_l},$$

corresponding to a lattice-periodic displacement modulated by a plane wave with wavevector $\boldsymbol{q}$, one can define the *dynamical matrix*

$$D_{st}^{\alpha\beta}(\boldsymbol{q}) := \frac{\tilde{C}_{st}^{\alpha\beta}(\boldsymbol{q})}{\sqrt{M_s M_t}}.$$

Phonon frequencies $\omega(\boldsymbol{q})$ and eigenmodes can be determined by diagonalizing the dynamical matrix, solving the secular equation

$$\det\left|D_{st}^{\alpha\beta}(\boldsymbol{q}) - \omega^2(\boldsymbol{q})\right| = 0.$$

1.2.2. Density Functional Perturbation Theory in Insulators

The open question at this point is how to actually calculate the linear response of the ground state density $\frac{\partial n_{0\underline{\boldsymbol{R}}}(\boldsymbol{r})}{\partial R_J}$ to nuclear displacement. The method employed in the context of our work is called *density functional perturbation theory* [57–60], which we briefly describe in the following. Assuming a Kohn-Sham system with single-particle orbitals only degenerate in the spin degree of freedom, the ground-state electronic density (1.5) for a given nuclear configuration $\underline{\boldsymbol{R}}$ can be expressed as

$$n(\boldsymbol{r}) = 2\sum_n^{N/2} |\varphi_n(\boldsymbol{r})|^2$$

$$\text{with } n(\boldsymbol{r}) := n_{0\,\underline{\boldsymbol{R}}}(\boldsymbol{r}) \text{ and } \varphi_n(\boldsymbol{r}) := \varphi_{\underline{\boldsymbol{R}}\,n}(\boldsymbol{r})$$

The linear response of the density under perturbation of the nuclear positions reads

$$\Delta_{\underline{\boldsymbol{R}}} n(\boldsymbol{r}) = 4\mathrm{Re}\sum_n^{N/2} \varphi_n^*(\boldsymbol{r})\Delta_{\underline{\boldsymbol{R}}}\varphi_n(\boldsymbol{r}) \tag{1.20}$$

where the operator

$$\Delta_{\underline{\boldsymbol{R}}} \bullet := \sum_{I,\alpha} \left(R_{I\alpha} - R_{I\alpha}^{(0)} \right) \left. \frac{\partial \bullet}{\partial R_{I\alpha}} \right|_{\underline{\boldsymbol{R}}=\underline{\boldsymbol{R}}^{(0)}}$$

is a finite-difference operator, describing the linearized change of a quantity when the nuclear geometry changes from $\underline{\boldsymbol{R}}^{(0)}$ (the equilibrium geometry) to $\underline{\boldsymbol{R}}$. The same convention of $\bullet^{(0)}$ referring to the initial nuclear configuration will be used throughout this chapter.

We treat the effect of the nuclear displacement on the Kohn-Sham states within first-order perturbation theory:

$$\left(\widehat{H}_{\mathrm{s}}^{(0)} - \epsilon_n^{(0)}\right) |\Delta_{\underline{\boldsymbol{R}}} \varphi_n\rangle = \left(\Delta_{\underline{\boldsymbol{R}}} \widehat{H}_{\mathrm{s}} - \Delta_{\underline{\boldsymbol{R}}} \epsilon_n\right) |\varphi_n\rangle = -\left(\Delta_{\underline{\boldsymbol{R}}} v_{\mathrm{s}} - \Delta_{\underline{\boldsymbol{R}}} \epsilon_n\right) |\varphi_n\rangle \tag{1.21}$$

with the first-order perturbation to the Kohn-Sham potential (1.10)

$$\begin{aligned} \Delta_{\underline{\boldsymbol{R}}} v_s(\boldsymbol{r}) = \Delta_{\underline{\boldsymbol{R}}} v^{\mathrm{en}}(\boldsymbol{r}) + e^2 \int \frac{\Delta_{\underline{\boldsymbol{R}}} n(\boldsymbol{r'})}{|\boldsymbol{r} - \boldsymbol{r'}|} \mathrm{d}^3 \boldsymbol{r'} \\ + \left. \frac{\mathrm{d} v_{\mathrm{xc}}(\boldsymbol{r})}{\mathrm{d}n} \right|_{n=n^{(0)}(\boldsymbol{r})} \Delta_{\underline{\boldsymbol{R}}} n(\boldsymbol{r}) \end{aligned} \tag{1.22}$$

and the first-order variation of the Kohn-Sham eigenvalues

$$\Delta_{\underline{\boldsymbol{R}}} \epsilon_n = \left\langle \varphi_n^{(0)} \middle| \Delta_{\underline{\boldsymbol{R}}} v_s \middle| \varphi_n^{(0)} \right\rangle . \tag{1.23}$$

The system of equantions (1.20), (1.21), (1.22) and (1.23) then can be solved iteratively by a self-consistent method, simultaneously for the responses $\Delta_{\underline{\boldsymbol{R}}} v_s(\boldsymbol{r})$, $\Delta_{\underline{\boldsymbol{R}}} \varphi_n(\boldsymbol{r})$, $\Delta_{\underline{\boldsymbol{R}}} n(\boldsymbol{r})$ and $\Delta_{\underline{\boldsymbol{R}}} \epsilon_n$, when the unperturbed states, eigenvalues and density is known.

Furthermore, in preparation for the coming steps, an alternate form of the equations can be derived from a conventional first-order expansion for a state perturbation in the basis of the unperturbed states

$$\Delta_{\underline{\boldsymbol{R}}} |\varphi_n\rangle = \sum_{m \neq n}^{\infty} \frac{\left\langle \varphi_m^{(0)} \middle| \Delta_{\underline{\boldsymbol{R}}} v_{\mathrm{s}} \middle| \varphi_n^{(0)} \right\rangle}{\epsilon_n^{(0)} - \epsilon_m^{(0)}} \left| \varphi_m^{(0)} \right\rangle \tag{1.24}$$

which implies a variation of the charge density (1.20)

$$\Delta_{\underline{\boldsymbol{R}}} n(\boldsymbol{r}) = 4 \sum_{n=1}^{N/2} \sum_{m \neq n}^{\infty} \varphi_n^{(0)\,*}(\boldsymbol{r}) \varphi_m^{(0)}(\boldsymbol{r}) \frac{\left\langle \varphi_m^{(0)} \middle| \Delta_{\underline{\boldsymbol{R}}} v_{\mathrm{s}} \middle| \varphi_n^{(0)} \right\rangle}{\epsilon_n^{(0)} - \epsilon_m^{(0)}}.$$

Terms with $n, m \leq N/2$, i.e. those where both states belong to the valence manyfold, do occur twice with opposite signs in the double summation, and therefore cancel. The remaining terms are those, where state $|\varphi_m\rangle$ belongs to the conduction manyfold:

$$\Delta_{\underline{\boldsymbol{R}}} n(\boldsymbol{r}) = 4 \sum_{n=1}^{N/2} \sum_{m > N/2} \varphi_n^{(0)\,*}(\boldsymbol{r}) \varphi_m^{(0)}(\boldsymbol{r}) \frac{\left\langle \varphi_m^{(0)} \middle| \Delta_{\underline{\boldsymbol{R}}} v_{\mathrm{s}} \middle| \varphi_n^{(0)} \right\rangle}{\epsilon_n^{(0)} - \epsilon_m^{(0)}}. \tag{1.25}$$

We now define two projection operators, one on the conduction and one on the valence manyfold of unperturbed Kohn-Sham states; their summation corresponds to unity, as the set of all Kohn-Sham wave functions are an orthonormal basis:

$$\mathbb{1} = \widehat{P}_c^{(0)} + \widehat{P}_v^{(0)} \quad | \quad \widehat{P}_c^{(0)} = \sum_{c'>N/2} \left|\varphi_{c'}^{(0)}\right\rangle\left\langle\varphi_{c'}^{(0)}\right|, \quad \widehat{P}_v^{(0)} = \sum_{v'=1}^{N/2} \left|\varphi_{v'}^{(0)}\right\rangle\left\langle\varphi_{v'}^{(0)}\right|$$

which we insert into the right hand side of (1.21). When we restrict our analysis to valence states φ_v, the result reads:

$$\begin{aligned}\left(\widehat{H}_{\mathrm{s}}^{(0)} - \epsilon_n^{(0)}\right)\left|\Delta_{\underline{\boldsymbol{R}}}\varphi_v\right\rangle &= -\left(\widehat{P}_c^{(0)} + \widehat{P}_v^{(0)}\right)\left(\Delta_{\underline{\boldsymbol{R}}} v_{\mathrm{s}} - \Delta_{\underline{\boldsymbol{R}}}\epsilon_n\right)\left|\varphi_v^{(0)}\right\rangle \\ &= -\widehat{P}_c^{(0)}\Delta_{\underline{\boldsymbol{R}}} v_{\mathrm{s}}\left|\varphi_v^{(0)}\right\rangle - \widehat{P}_v^{(0)}\left(\Delta_{\underline{\boldsymbol{R}}} v_{\mathrm{s}} - \Delta_{\underline{\boldsymbol{R}}}\epsilon_n\right)\left|\varphi_v^{(0)}\right\rangle.\end{aligned}$$

In Ref [57], the complete right-hand side term involving $\widehat{P}_v^{(0)}$ is now neglected, based on the observation stated in (1.25), which relates responses in the charge density only to terms involving both conduction and valence states. Moreover, in order to remove the null eigenvalue on the left hand side, a small term $\alpha\widehat{P}_v^{(0)}$ is introduced, making the operator nonsingular, while, by acting only on occupied states, having no effect on the density response $\Delta_{\underline{\boldsymbol{R}}} n(\boldsymbol{r})$:

$$\left(\widehat{H}_{\mathrm{s}}^{(0)} - \alpha\widehat{P}_v^{(0)} - \epsilon_n^{(0)}\right)\left|\Delta_{\underline{\boldsymbol{R}}}\varphi_v\right\rangle = -\widehat{P}_c^{(0)}\Delta_{\underline{\boldsymbol{R}}} v_{\mathrm{s}}\left|\varphi_v^{(0)}\right\rangle. \tag{1.26}$$

Any perturbation can be expressed by its Fourier representation,

$$\Delta_{\underline{\boldsymbol{R}}} v_{\mathrm{s}}(\boldsymbol{r}) = \sum_{\boldsymbol{q}} \Delta_{\underline{\boldsymbol{R}}}^{\boldsymbol{q}} v_{\mathrm{s}}(\boldsymbol{r}) \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}$$

where each element represents a perturbation with a single wave vector $\boldsymbol{q}$, termed as a *monochromatic perturbation*, while each individual component $\Delta_{\underline{\boldsymbol{R}}}^{\boldsymbol{q}} v_{\mathrm{s}}(\boldsymbol{r})$ shares the full periodicity of the original lattice. Assuming Kohn-Sham states with Bloch form $\varphi_{n\boldsymbol{k}}(\boldsymbol{r}) = u_{n\boldsymbol{k}}(\boldsymbol{r})\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}$ (1.14), and projecting on the manifold of all states with wave vectors $\boldsymbol{k}+\boldsymbol{q}$, (1.26) can be transformed to

$$\begin{aligned}&\left(\widehat{H}_s^{(0)\,\boldsymbol{k}+\boldsymbol{q}} + \alpha\sum_{v'}\left|u_{v'\,\boldsymbol{k}+\boldsymbol{q}}^{(0)}\right\rangle\left\langle u_{v'\,\boldsymbol{k}+\boldsymbol{q}}^{(0)}\right| - \epsilon_{v\,\boldsymbol{k}}^{(0)}\right)\left|\Delta_{\underline{\boldsymbol{R}}} u_{v\,\boldsymbol{k}+\boldsymbol{q}}\right\rangle \\ &\qquad = -\left(1 - \sum_{v'}\left|u_{v'\,\boldsymbol{k}+\boldsymbol{q}}^{(0)}\right\rangle\left\langle u_{v'\,\boldsymbol{k}+\boldsymbol{q}}^{(0)}\right|\right)\Delta_{\underline{\boldsymbol{R}}}^{\boldsymbol{q}} v_{\mathrm{s}}\left|u_{v\,\boldsymbol{k}}^{(0)}\right\rangle \qquad (1.27) \\ &\qquad\qquad \text{where } \widehat{H}_s^{(0)\,\boldsymbol{k}+\boldsymbol{q}} := \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}\widehat{H}_s^{(0)}\mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}.\end{aligned}$$

To complete the system of equations to be solved, the density and potential response to such monochromatic perturbations are determined as

$$\Delta^{\boldsymbol{q}}_{\underline{\boldsymbol{R}}} n(\boldsymbol{r}) = 4 \sum_{v\,\boldsymbol{k}} u^{(0)*}_{v\,\boldsymbol{k}}(\boldsymbol{r}) \Delta_{\underline{\boldsymbol{R}}} u_{v\,\boldsymbol{k}+\boldsymbol{q}} \tag{1.28}$$

$$\begin{aligned}\Delta^{\boldsymbol{q}}_{\underline{\boldsymbol{R}}} v_{\mathrm{s}}(\boldsymbol{r}) = \Delta^{\boldsymbol{q}}_{\underline{\boldsymbol{R}}} v^{\mathrm{en}}(\boldsymbol{r}) + e^2 \int \frac{\Delta^{\boldsymbol{q}}_{\underline{\boldsymbol{R}}} n(\boldsymbol{r'})}{|\boldsymbol{r} - \boldsymbol{r'}|} \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot(\boldsymbol{r}-\boldsymbol{r'})} \mathrm{d}^3\boldsymbol{r'} \\ + \left.\frac{\mathrm{d}v_{\mathrm{xc}}(\boldsymbol{r})}{\mathrm{d}n}\right|_{n=n^{(0)}(\boldsymbol{r})} \Delta^{\boldsymbol{q}}_{\underline{\boldsymbol{R}}} n(\boldsymbol{r}).\end{aligned} \tag{1.29}$$

All quantities appearing in (1.27), (1.28) and (1.29) share the full periodicity of the lattice, therefore this system of equations can be solved just within the original unit cell for each individual $\boldsymbol{q}$.

This fact yields a large advantage over the conventional frozen-phonon technique: phonons with arbitrary wave vectors $\boldsymbol{q}$ share roughly the same computational complexity as self-consistent Kohn-Sham calculations within the original cell, while the latter requires the construction of supercells having $\boldsymbol{q}$ as a reciprocal lattice vector, which results in calculations of mostly infeasible complexity for all but a few zone-center and zone-boundary phonon modes.

1.2.3. Extension of DFPT to metals

In metals, the smearing technique, as explained in section 1.1.3, is applied in order to obtain converged results with a finite number of $\boldsymbol{k}$ points, leading to fractional occupation of Kohn-Sham states around the Fermi Level. The formerly explained DFPT scheme, however, was derived considering a clear separation of valence states (corresponding to an occupation of 1) and conduction states (corresponding to an occupation of 0). As a first step to extend DFPT to metals, the density response is adapted to the presence of a smearing function (1.17):

$$\begin{aligned}\Delta_{\underline{\boldsymbol{R}}} n(\boldsymbol{r}) = \quad & \sum_n \theta^{\sigma}_{\mathrm{F},n} \left(\varphi^{(0)}_n * (\boldsymbol{r}) \Delta_{\underline{\boldsymbol{R}}} \varphi_n(\boldsymbol{r}) + c.c. \right) \\ & + \sum_n \left| \varphi^{(0)}_n(\boldsymbol{r}) \right|^2 S^{\sigma}_{\mathrm{F},n} \left(\Delta_{\underline{\boldsymbol{R}}} \epsilon_F - \Delta_{\underline{\boldsymbol{R}}} \epsilon_n \right),\end{aligned} \tag{1.30}$$

where the first term originates from the linear response of the Kohn-Sham states, while the second term is related to a possible response in the single particle energies ϵ_n or the Fermi level ϵ_F. A short-hand notation for the smearing functions has been introduced as

$$S^{\sigma}_{n,m} := S_{\sigma}(\epsilon^{(0)}_n - \epsilon^{(0)}_m) \;\text{ and }\; \theta^{\sigma}_{n,m} := \theta_S \left(\frac{\epsilon^{(0)}_n - \epsilon^{(0)}_m}{\sigma} \right).$$

Using the first-order perturbation expansion (1.24) for the response of the KS state, the metallic density response can be rewritten as

$$\Delta_{\underline{\boldsymbol{R}}} n(\boldsymbol{r}) = \sum_{n,m} \frac{\theta^{\sigma}_{\mathrm{F},n} - \theta^{\sigma}_{\mathrm{F},m}}{\epsilon^{(0)}_n - \epsilon^{(0)}_m} \varphi^{(0)\,*}_n(\boldsymbol{r}) \varphi^{(0)}_m(\boldsymbol{r}) \left\langle \varphi^{(0)}_m \middle| \Delta_{\underline{\boldsymbol{R}}} v_{\mathrm{s}} \middle| \varphi^{(0)}_n \right\rangle ,$$

where the $m = n$ term accounts for the $\Delta_{\underline{\boldsymbol{R}}} \epsilon_{\mathrm{F}}$ term in (1.30). Moreover, for any cases when $\epsilon_n \to \epsilon_m$ the prefactor is to be substituted by its analytic limit

$$-S^{\sigma}_{\mathrm{F},n} = \lim_{\epsilon^{(0)}_n \to \epsilon^{(0)}_m} \frac{\theta^{\sigma}_{\mathrm{F},n} - \theta^{\sigma}_{\mathrm{F},m}}{\epsilon^{(0)}_n - \epsilon^{(0)}_m},$$

which is a broadened δ function for finite smearing (and would diverge without smearing). Note that, depending on the topology of the Fermi surface, *nesting vectors* $\boldsymbol{q}_{\mathrm{n}}$ may exist, which connect finite, parellel sheets of the surface. Phonons with corresponding wave vectors lead to strong responses in the density, greatly reducing such a phonon's frequency, an effect which is called a *Kohn anomaly.*

1.2.4. Electron-phonon interaction

While all statements up to this point are completely general and apply also to molecular dynamics, a change of coordinates and some algebra [61] allows to rewrite the nuclear Born-Oppenheimer Hamiltonian (1.18) in harmonic oscillator form

$$\widehat{H}^{\mathrm{ph}} = \sum_{\nu,\boldsymbol{q}} \omega_{\nu\,\boldsymbol{q}} \left(\widehat{b}^{\dagger}_{\nu\,\boldsymbol{q}} \widehat{b}_{\nu\,\boldsymbol{q}} + \tfrac{1}{2} \right),$$

where the operators $\widehat{b}^{\dagger}_{\nu\,\boldsymbol{q}}$ and $\widehat{b}_{\nu\,\boldsymbol{q}}$ are creation/annihilation operators of collective vibrational excitations of the ionic lattice, *phonons* identified by wave vector $\boldsymbol{q}$ and *branch index* ν.

The energy $\hbar\omega_{\nu\,\boldsymbol{q}}$ carried by a single phonon lies in the range of 10meV to 100meV, at least an order of magnitude lower than typical electronic transition energies; this difference in scale is the reformulated justification for the assumption of adiabaticity within the Born-Oppenheimer approximation, implying the absence of electronic transitions caused by nuclear motion.

In metals, however, electronic transistions may be induced by scattering processes involving phonons within a small energy window around Fermi level. Such processes have a significant contribution to normal-state resistance and, even more important in the context of this work, give rise to the phenomenon of conventional superconductivity.

In order to establish a perturbative treatement of the phonons' effect in the electronic problem, the *electron-phonon Hamiltonian*, coupling ionic and electronic Born-Oppenheimer equations, is established as

$$\widehat{H}^{\mathrm{el-ph}} := \sum_{\boldsymbol{k},n,m,\boldsymbol{q},\nu} g^{\nu,n,m}_{\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q}} \left(\widehat{b}_{\boldsymbol{q}\,\nu} - \widehat{b}^{\dagger}_{-\boldsymbol{q}\,\nu} \right) \widehat{a}^{\dagger}_{\boldsymbol{k}+\boldsymbol{q}\,m} \widehat{a}_{\boldsymbol{k}\,n}$$

where $\widehat{a}_{\boldsymbol{k}\,n}$ is the annihilation operator of an electron in state $\boldsymbol{k}\,n$, while $\widehat{b}_{\boldsymbol{q}\,\nu}$ annihilates a phonon in mode $\boldsymbol{q}\,\nu$. Effectively, this Hamiltonian encompasses all electronic transitions involving the absorption or emission of a single phonon, with appropriate crystal momentum transfer to ensure the conservation of total crystal momentum. The matrix element $g^{\nu,n,m}_{\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q}}$ relates to the strength of the electron-phonon interaction, and can be interpreted as a probability amplitude of the corresponding process.

Within the framework of density functional perturbation theory, a definition of the matrix elements has been proposed in [62, 63]:

$$g^{\nu,n,m}_{\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q}} := \sqrt{\frac{\hbar}{2\omega_{\boldsymbol{q}\,\nu}}} \left\langle \varphi_{\boldsymbol{k}+\boldsymbol{q}\,m} \middle| \Delta^{\boldsymbol{q}\,\nu}_{\underline{\boldsymbol{R}}} v_s \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \middle| \varphi_{\boldsymbol{k}\,n} \right\rangle , \tag{1.31}$$

treating the response of the Kohn-Sham potential as a perturbation of the electronic subsystem by a phonon with wave vector $\boldsymbol{q}$ and mode index ν, dressed by adiabatically moving charge; the prefactor accounts for the amplitude of the oscillation, scaling with the phonon numbers.

Extensive comparison of the results in [63] shows good agreement with experiment, which confirms the presented ansatz.

1.2.5. Summary

Density functional perturbation theory provides a method to evaluate phononic properties, such as modes, frequencies and the interaction between electrons and phonons fully *ab initio*. Due to the decomposition into monochromatic perturbations, the non-lattice-periodic part of the lattice deformation can be accounted for by a Bloch-like phase factor, while the problem solved for each wave vector can be solved independently within the primitive unit cell, at a computational cost, per mode, comparable [57] to a self-consistent Kohn-Sham calculation. This fact is the key advantage of this method, as the competing scheme, the frozen-phonon method, relies on the construction of supercells commensurate with the wave vector, making it probhibitively expensive to apply to any but a few selected high-symmetry modes.

The computed spectra and electron-phonon interaction measures are in excellent agreement with experimentally obtained ones, making the method an integral part of computational methods for the prediction of superconducting properties.

In the next section, we review such a method, and its theoretical foundation.

2. Theory of the Superconducting State

From a phenomenological point of view, superconductivity is a phase existing below a certain, material-dependent *critical temperature* T_c, which is characterized by the absence of any electrical resistivity and the Meissner-Ochsenfeld-effect. The latter describes perfect diamagnetism, i.e. the expulsion of any magnetic field from the superconducting sample.

Within this chapter, we present an overview on the theory of superconductivity, starting with a brief introduction of BCS theory, which was the first to provide a microscopical understanding of superconductivity. As BCS lacks predictive power regarding material-specific properties in the presence of strong electron-phonon coupling, Eliashberg theory and its McMillan approximation are introduced in the central part. This chapter closes with a review on the relation between superconductivity, the electron-phonon coupling and material-specific properties.

2.1. BCS theory

Bardeen, Cooper and Schrieffer (BCS) provided the first microscopic theory [45, 46] explaining the nature of the superconducting phase transition: a condensation of electrons with opposite spin and crystal momenta $\boldsymbol{k}$ into Cooper pairs [64] occurs by an effectively *attractive electron-electron interaction* [65] mediated by collective motion of the nucleii, i.e. via the electron-phonon interaction (subsection 1.2.4). In BCS theory, the system is described by an effective Hamiltonian

$$\widehat{H}_\mathrm{BCS} := \sum_{\boldsymbol{k}\sigma} \epsilon_{\boldsymbol{k}} \widehat{c}^\dagger_{\boldsymbol{k}\sigma} \widehat{c}_{\boldsymbol{k}\sigma} + \sum_{\boldsymbol{k}\boldsymbol{k}'} V_{\boldsymbol{k}\boldsymbol{k}'} \widehat{c}^\dagger_{\boldsymbol{k}\uparrow} \widehat{c}^\dagger_{-\boldsymbol{k}\downarrow} \widehat{c}_{-\boldsymbol{k}'\downarrow} \widehat{c}_{-\boldsymbol{k}'\uparrow} \tag{2.1}$$

where the $\epsilon_{\boldsymbol{k}}$ are the normal-state eigenvalues, the operators $\widehat{c}_{\boldsymbol{k}\sigma}$ are annihilation operators of Bloch electrons of wave vector $\boldsymbol{k}$, spin σ and $V_{\boldsymbol{k}\boldsymbol{k}'}$ is the effective electron-electron potential due to the phonon-mediated interaction, which is effective only within the (energetic) vincinity of the Fermi surface; this consideration is justified by both the, compared to electronic energies, small energy provided and absorbed by phonons in scattering processes and the Pauli principle preventing already occupied scattering target states. In the case of a system with only one band crossing Fermi level, BCS propose an approximation of this potential by a single-well potential

$$V_{\boldsymbol{k}\boldsymbol{k}'} = \begin{cases} -V & \text{if } |\epsilon_{\boldsymbol{k}} - E_\mathrm{F}|, |\epsilon_{\boldsymbol{k}'} - E_\mathrm{F}| < \hbar\omega_\mathrm{D} \\ 0 & \text{otherwise} \end{cases} , \tag{2.2}$$

where ω_D is the Debye frequency (representative frequency of phonons within a material). BCS theory introduces an electronic wave function

$$|\Psi_\mathrm{BCS}\rangle := \prod_{\boldsymbol{k}} \left(u_{\boldsymbol{k}} + v_{\boldsymbol{k}} \widehat{c}^\dagger_{\boldsymbol{k}\uparrow} \widehat{c}^\dagger_{-\boldsymbol{k}\downarrow}\right) |0\rangle \quad \text{normalization } |u_{\boldsymbol{k}}|^2 + |v_{\boldsymbol{k}}|^2 := 1, \tag{2.3}$$

$u_{\boldsymbol{k}}, v_{\boldsymbol{k}} \in \mathbb{C}$, to describe the superconducting system. In the limit of

$$(v_{\boldsymbol{k}}, u_{\boldsymbol{k}}) = \begin{cases} (1,0) & \text{for } \epsilon_{\boldsymbol{k}} < E_\mathrm{F} \\ (0,1) & \text{otherwise} \end{cases},$$

the wave function recovers the non-superconducting ground state. However, for any other cases, the number of particles within (2.3) is not fixed, the wave function can be decomposed into terms with different particle numbers. Problems in the interpretation vanish when taking the macroscopic limit, particle number $N \to \infty$, applicable to a bulk system, as on this scale *fractional* fluctuations go to zero. In result, the requirement of a specific particle number can be only satisfied on average, by means of a chemical potential $\mu \equiv E_\mathrm{F}$, introduced as a Lagrangian multiplier into the variational problem, stationary when

$$0 \equiv \delta \left\langle \Psi_\mathrm{BCS} \middle| \widehat{H} - \mu \widehat{N} \middle| \Psi_\mathrm{BCS} \right\rangle \tag{2.4}$$

$$\text{where } \widehat{H}_\mathrm{BCS} - \mu \widehat{N} = \sum_{\boldsymbol{k}\sigma} \xi_{\boldsymbol{k}} \widehat{c}^\dagger_{\boldsymbol{k}\sigma} \widehat{c}_{\boldsymbol{k}\sigma} + \sum_{\boldsymbol{k}\boldsymbol{k}'} V_{\boldsymbol{k}\boldsymbol{k}'} \widehat{c}^\dagger_{\boldsymbol{k}\uparrow} \widehat{c}^\dagger_{-\boldsymbol{k}\downarrow} \widehat{c}_{-\boldsymbol{k}'\downarrow} \widehat{c}_{-\boldsymbol{k}'\uparrow} \tag{2.5}$$

$$\text{with } \xi_{\boldsymbol{k}} := \epsilon_{\boldsymbol{k}} - \mu. \tag{2.6}$$

Using BCS form (2.3) of the wave function, (2.4) becomes

$$0 \equiv 2\xi_{\boldsymbol{k}} u_{\boldsymbol{k}} v_{\boldsymbol{k}} + \Delta_{\boldsymbol{k}} (v_{\boldsymbol{k}}^2 - u_{\boldsymbol{k}}^2), \tag{2.7}$$

defining the *gap function* as

$$\Delta_{\boldsymbol{k}} := \sum_{\boldsymbol{k}'} V_{\boldsymbol{k}\boldsymbol{k}'} \left\langle \Psi_\mathrm{BCS} \middle| \widehat{c}_{\boldsymbol{k}'\uparrow} \widehat{c}_{-\boldsymbol{k}'\downarrow} \middle| \Psi_\mathrm{BCS} \right\rangle = \sum_{\boldsymbol{k}'} V_{\boldsymbol{k}\boldsymbol{k}'} u_{\boldsymbol{k}'} v_{\boldsymbol{k}'}. \tag{2.8}$$

This stationarity condition is fulfilled for

$$u_{\boldsymbol{k}}^2 = \frac{1}{2}\left(1 + \frac{\xi_{\boldsymbol{k}}}{E_{\boldsymbol{k}}}\right) \text{ and } v_{\boldsymbol{k}}^2 = \frac{1}{2}\left(1 - \frac{\xi_{\boldsymbol{k}}}{E_{\boldsymbol{k}}}\right), \text{ where } E_{\boldsymbol{k}} := \sqrt{\xi_{\boldsymbol{k}}^2 + \Delta_{\boldsymbol{k}}^2}. \tag{2.9}$$

Substituting (2.9) in (2.8) yields the *BCS self-consistent gap equation*

$$\Delta_{\boldsymbol{k}} = -\frac{1}{2} \sum_{\boldsymbol{k}'} \frac{V_{\boldsymbol{k}\boldsymbol{k}'}}{E_{\boldsymbol{k}'}} \Delta_{\boldsymbol{k}'} \tag{2.10}$$

at zero temperature.

The interpretation of $\Delta_{\boldsymbol{k}}$ as a gap can be estblished by considering excitations on top of the BCS ground state, via the Boguliubov transformation [66]

$$\widehat{c}_{\boldsymbol{k}\uparrow} = u_{\boldsymbol{k}} \widehat{\gamma}_{\boldsymbol{k}\uparrow} + v_{\boldsymbol{k}} \widehat{\gamma}^\dagger_{-\boldsymbol{k}\downarrow} \qquad \widehat{c}_{\boldsymbol{k}\downarrow} = u_{\boldsymbol{k}} \widehat{\gamma}_{\boldsymbol{k}\downarrow} + v_{\boldsymbol{k}} \widehat{\gamma}^\dagger_{-\boldsymbol{k}\uparrow} \tag{2.11}$$

resulting in the Hamiltonian

$$\widehat{H} = \sum_{\boldsymbol{k}} E_{\boldsymbol{k}} \left(\widehat{\gamma}^{\dagger}_{\boldsymbol{k}\uparrow} \widehat{\gamma}_{\boldsymbol{k}\uparrow} + \widehat{\gamma}^{\dagger}_{\boldsymbol{k}\downarrow} \widehat{\gamma}_{\boldsymbol{k}\downarrow} \right) + \sum_{\boldsymbol{k}} \left(\xi_{\boldsymbol{k}} - E_{\boldsymbol{k}} + \Delta_{\boldsymbol{k}} \widehat{c}_{\boldsymbol{k}\uparrow} \widehat{c}_{-\boldsymbol{k}\downarrow} \right).$$

The $E_{\boldsymbol{k}}$ are evidently excitation energies, and considering electrons with $\xi_{\boldsymbol{k}} = 0$, there exists a lower bound $\Delta_{\boldsymbol{k}}$ (cf. Equation 2.9). Therefore, the spectrum of a superconductor exhibits a gap around μ, and due to the fact that in absorbition or emission processes the state of a *pair* of electrons is altered, the width of this gap is 2Δ.

Furthermore, excitations are responsible for the finite-temperature properties of superconductors: with increasing temperature, their population increases following the Fermi distribution

$$f_{\boldsymbol{k}} := \left(1 + \mathrm{e}^{\frac{E_{\boldsymbol{k}}}{k_{\mathrm{B}} T}} \right)^{-1}.$$

In turn the interaction energy of the remaining pairs is lowered, leading to the finite-temperature gap equation

$$\Delta_{\boldsymbol{k}} = -\frac{1}{2} \sum_{\boldsymbol{k}'} \frac{V_{\boldsymbol{k}\boldsymbol{k}'} \Delta_{\boldsymbol{k}'}}{E_{\boldsymbol{k}'}} \left(1 - 2 f_{\boldsymbol{k}'} \right) = -\frac{1}{2} \sum_{\boldsymbol{k}'} \frac{V_{\boldsymbol{k}\boldsymbol{k}'} \Delta_{\boldsymbol{k}'}}{E_{\boldsymbol{k}'}} \tanh \left(\frac{E_{\boldsymbol{k}'}}{2 k_{\mathrm{B}} T} \right).$$

In the case of the simple BCS model potential (2.2), a simple solution for the critical temperature T_{c} can be found:

$$k_{\mathrm{B}} T_c = 1.14 \hbar \omega_{\mathrm{D}} \mathrm{e}^{-\frac{1}{\lambda}} \text{ where } \lambda = N_{E_{\mathrm{F}}} V \tag{2.12}$$

is the electron-phonon coupling parameter with the density of states per spin at Fermi level $N_{E_{\mathrm{F}}}$.

2.2. Eliashberg Theory of the Superconducting State

Eliashberg theory [47] is the most widely applied method for quantitative predictions of material-specific properties.

While BCS theory explained the superconducting phase transition, and even allowed for the prediction of up to this point unknown characteristica of superconductors, e.g. the Josephson effect [67]. However, the accuracy of *quantitative* predictions for material-specific properties such as T_{c} is limited: for one, the simple form of the Fröhlich interaction in the effective Hamiltonian (2.1) was built on the assumption of a weak electron-phonon interaction, and for the other, effects of the direct Coulomb interaction beyond the effective mass entering via the normal-state eigenvalues $\epsilon_{\boldsymbol{k}}$ are neglected. While many-body perturbation theory (MBPT) as such provides a powerful tool for a systematic approximate treatment of electronic properties including strong-coupling normal-state electron-phonon effects on the electronic self energy [68], MBPT is rendered invalid by the superconducting phase transition.

In Eliashberg theory, applicability of MBPT is restored by the means of Nambu-Gor'kov formalism [69–71]. The Migdal theorem [68], originally developed for the normal

phase, is then invoked within this extended MBPT for the superconducting phase: it states that single-phonon scattering is by far the dominant contribution to the electronic self-energy; higher order terms provide corrections in the order of $\omega_\mathrm{D}/E_\mathrm{F}$ (typically 10^{-2}) and can therefore be neglected. While the orignal theory [47] did not include the repulsive, direct Coulomb interaction, it was included by an extension [72] as the effective Morel-Anderson pseudopotential μ^*; the latter is not easily accessible from first principles, and is therefore usually treated as a parameter to be fitted to the experiment.

2.2.1. Nambu formalism and Greens function for superconductors

Nambu-Gor'kov formalism introduces the field operators

$$\widehat{\boldsymbol{\Psi}}_{\boldsymbol{k}} := \begin{pmatrix} \widehat{c}_{\boldsymbol{k}\uparrow} \\ \widehat{c}^\dagger_{-\boldsymbol{k}\downarrow} \end{pmatrix} \qquad \widehat{\varphi}_{\boldsymbol{q}\nu} := \left(\widehat{b}_{\boldsymbol{q}\nu} + \widehat{b}^\dagger_{-\boldsymbol{q}\nu} \right) \tag{2.13}$$

for electrons (Nambu spinor) and phonons (linear combination of phonon and therefore bosonic creation and annihilation operators), respectively. Using these field operators, Eliashberg theory considers a Hamiltonian of the form

$$\begin{aligned} \widehat{H}_\mathrm{E} := \quad & \sum_{\boldsymbol{k}} \xi_{\boldsymbol{k}} \widehat{\boldsymbol{\Psi}}^\dagger_{\boldsymbol{k}} \widehat{\sigma}_3 \widehat{\boldsymbol{\Psi}}_{\boldsymbol{k}} + \sum_{\boldsymbol{q}\nu} \omega_{\boldsymbol{q}\nu} \widehat{\varphi}^\dagger_{-\boldsymbol{q}\nu} \widehat{\varphi}_{\boldsymbol{q}\nu} \\ & + \sum_{\boldsymbol{k}\boldsymbol{k}'\nu} g_{\boldsymbol{k}\boldsymbol{k}'\nu} \widehat{\varphi}_{(\boldsymbol{k}-\boldsymbol{k}')\nu} \widehat{\boldsymbol{\Psi}}^\dagger_{\boldsymbol{k}'} \widehat{\sigma}_3 \widehat{\boldsymbol{\Psi}}_{\boldsymbol{k}} \\ & + \frac{1}{2} \sum_{\boldsymbol{k}_1\boldsymbol{k}_2\boldsymbol{k}_3\boldsymbol{k}_4} \langle \boldsymbol{k}_3 \boldsymbol{k}_4 | V_\mathrm{c} | \boldsymbol{k}_1 \boldsymbol{k}_2 \rangle \left(\widehat{\boldsymbol{\Psi}}^\dagger_{\boldsymbol{k}_3} \widehat{\sigma}_3 \widehat{\boldsymbol{\Psi}}_{\boldsymbol{k}_1} \right) \left(\widehat{\boldsymbol{\Psi}}^\dagger_{\boldsymbol{k}_4} \widehat{\sigma}_3 \widehat{\boldsymbol{\Psi}}_{\boldsymbol{k}_2} \right), \end{aligned}$$

where $\xi_{\boldsymbol{k}}$ are the one-electron energies relative to E_F (cf. (2.9)), $\omega_{\boldsymbol{q}\nu}$ are the bare phonon energies, $g_{\boldsymbol{k}\boldsymbol{k}'\nu}$ are electron-phonon matrix elements, V_c is the Coulomb potential and $\widehat{\sigma}_3$ is one of the Pauli matrices

$$\widehat{\sigma}_0 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \widehat{\sigma}_1 := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \widehat{\sigma}_2 := \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix} \qquad \widehat{\sigma}_3 := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.14}$$

The corresponding finite-temperature electronic Greens function is

$$\begin{aligned} G(\boldsymbol{k}\tau) := & - \left\langle \widehat{U}\widehat{T} \left\{ \widehat{\boldsymbol{\Psi}}_{\boldsymbol{k}}(\tau) \otimes \widehat{\boldsymbol{\Psi}}^\dagger_{\boldsymbol{k}}(0) \right\} \right\rangle \\ = & - \begin{pmatrix} \left\langle \widehat{T} \left\{ \widehat{c}_{\boldsymbol{k}\uparrow}(\tau) \widehat{c}^\dagger_{\boldsymbol{k}\uparrow}(0) \right\} \right\rangle & \left\langle \widehat{U}\widehat{T} \left\{ \widehat{c}_{\boldsymbol{k}\uparrow}(\tau) \widehat{c}_{-\boldsymbol{k}\downarrow}(0) \right\} \right\rangle \\ \left\langle \widehat{U}\widehat{T} \left\{ \widehat{c}^\dagger_{-\boldsymbol{k}\downarrow}(\tau) \widehat{c}^\dagger_{\boldsymbol{k}\uparrow}(0) \right\} \right\rangle & \left\langle \widehat{T} \left\{ \widehat{c}^\dagger_{-\boldsymbol{k}\downarrow}(\tau) \widehat{c}_{-\boldsymbol{k}\downarrow}(0) \right\} \right\rangle \end{pmatrix}, \end{aligned}$$

i.e. a 2×2 matrix, where the main diagonal contains the normal propagators for spin-up electrons (G_{11}) and spin-down holes (G_{22}), while the off-diagonal consists of the *anomalous propagators* (called F and $\bar{F}$ by Gor'kov) only finite in the superconducting state. $\widehat{T}$ is the usual time-ordering operator and $\widehat{U} := 1 + \widehat{R}^\dagger + \widehat{R}$ adjusts the particle

number (as the anomalous propagators do not conserve the latter): $\widehat{R}^\dagger$ converts an N-particle state into the corresponding $N+2$ particle state. The average denotes the grand-canonical ensemble average

$$\left\langle \widehat{Q} \right\rangle := \frac{\mathrm{Tr}(\mathrm{e}^{-\beta \widehat{H}_\mathrm{E}} \widehat{Q})}{\mathrm{Tr}(\mathrm{e}^{-\beta \widehat{H}_\mathrm{E}})}.$$

The phonon Greens function, on the other hand, is defined as

$$D_\nu(\boldsymbol{q}, \tau) := -\left\langle \widehat{T} \left\{ \widehat{\varphi}_{\boldsymbol{q}\nu}(\tau) \widehat{\varphi}^\dagger_{\boldsymbol{q}\nu}(0) \right\} \right\rangle$$

and electron and phonon field operators at imaginary time $\tau = \mathrm{i}t$[1]

$$\begin{aligned} \widehat{\boldsymbol{\Psi}}_{\boldsymbol{k}}(\tau) :=& \mathrm{e}^{\tau \widehat{H}_\mathrm{E}} \widehat{\boldsymbol{\Psi}}_{\boldsymbol{k}} \mathrm{e}^{-\tau \widehat{H}_\mathrm{E}} \\ \widehat{\varphi}_{\boldsymbol{q}\nu}(\tau) :=& \mathrm{e}^{\tau \widehat{H}_\mathrm{E}} \widehat{\varphi}_{\boldsymbol{q}\nu} \mathrm{e}^{-\tau \widehat{H}_\mathrm{E}}. \end{aligned}$$

Given that the imaginary time interval was limited to $-\beta < \tau < \beta$, a periodicity

$$G(\boldsymbol{k}, \tau) = G(\boldsymbol{k}, \tau + \beta)$$

is recovered from the cyclic property of the trace [73, p.301], allowing for a Fourier expansion in imaginary time

$$D_\nu(\boldsymbol{q}, \tau) = \tfrac{1}{\beta} \sum_{n=-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\nu_n \tau} D_\nu(\boldsymbol{q}, \mathrm{i}\nu_n) \tag{2.15}$$

$$G(\boldsymbol{k}, \tau) = \tfrac{1}{\beta} \sum_{n=-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega_n \tau} G(\boldsymbol{k}, \mathrm{i}\omega_n) \tag{2.16}$$

with the bosonic and fermionic *Matsubara frequencies*

$$\nu_n := 2n\tfrac{\pi}{\beta} \qquad \omega_n := (2n+1)\tfrac{\pi}{\beta}. \tag{2.17}$$

2.2.2. Self-energy and Gap

The MBPT approach taken by Eliashberg is based on the *Dyson equation* in matrix matrix form

$$G(\boldsymbol{k}, \mathrm{i}\omega_n) = G_0(\boldsymbol{k}, \mathrm{i}\omega_n) + G_0(\boldsymbol{k}, \mathrm{i}\omega_n) \Sigma(\boldsymbol{k}, \mathrm{i}\omega_n) G(\boldsymbol{k}, \mathrm{i}\omega_n) \tag{2.18}$$

determining the interacting electronic Greens function $G(\boldsymbol{k}, \mathrm{i}\omega_n)$ self-consistently in terms of the non-interacting Greens function

$$G_0(\boldsymbol{k}, \mathrm{i}\omega_n) := \frac{1}{\mathrm{i}\omega_n \widehat{\sigma}_0 - \xi_{\boldsymbol{k}} \widehat{\sigma}_0}$$

[1] imaginary time is introduced in the finite-temperature formalism, such that the time evolution operator $\mathrm{e}^{\tau \widehat{H}}$ and the Boltzmann factor $\mathrm{e}^{-\beta \widehat{H}}$ appearing in the grand-canonical averages are formally equal

and the electronic self-energy $\Sigma(\boldsymbol{k}, \mathrm{i}\omega_n)$, which is the quantity to be approximated in the following. Note that by multiplying with G^{-1} (from the right) and G_0^{-1} from the left, the Dyson equation can be reexpressed as

$$G^{-1}(\boldsymbol{k}, \mathrm{i}\omega_n) = G_0^{-1}(\boldsymbol{k}, \mathrm{i}\omega_n) - \Sigma(\boldsymbol{k}, \mathrm{i}\omega_n) \tag{2.19}$$

As a first step for the approximation, expansion of the self-energy in terms of the Pauli matrices (2.14)

$$\begin{aligned}\Sigma(\boldsymbol{k}, \mathrm{i}\omega_n) = &[1 - Z(\boldsymbol{k}, \mathrm{i}\omega_n)]\, \mathrm{i}\omega_n \widehat{\sigma}_0 + \ \chi(\boldsymbol{k}, \mathrm{i}\omega_n)\widehat{\sigma}_3 \\ &+ \phi_1(\boldsymbol{k}, \mathrm{i}\omega_n)\widehat{\sigma}_1 + \phi_2(\boldsymbol{k}, \mathrm{i}\omega_n)\widehat{\sigma}_2\end{aligned} \tag{2.20}$$

proves useful, where the $(\boldsymbol{k}, \mathrm{i}\omega_n)$-dependent expansion coefficients Z, χ, ϕ_1 and ϕ_2 are the unknowns to be determined later on. An important property can be already observed before assuming a specific Σ; inserting (2.20) into the Dyson equation (2.19), followed by an inversion, one obtains

$$\begin{aligned}G(\boldsymbol{k}, \mathrm{i}\omega_n) =& \frac{\phi_1(\boldsymbol{k}, \mathrm{i}\omega_n)\widehat{\sigma}_1 + \phi_2(\boldsymbol{k}, \mathrm{i}\omega_n)\widehat{\sigma}_2 + \mathrm{i}\omega_n Z(\boldsymbol{k}, \mathrm{i}\omega_n)\widehat{\sigma}_0 + (\xi_{\boldsymbol{k}} + \chi(\boldsymbol{k}, \mathrm{i}\omega_n))\widehat{\sigma}_3}{\Theta(\boldsymbol{k}, \mathrm{i}\omega_n)} \\ =& \frac{1}{\Theta(\boldsymbol{k}, \mathrm{i}\omega_n)} \begin{pmatrix} \mathrm{i}\omega_n Z(\boldsymbol{k}, \mathrm{i}\omega_n) + (\xi_{\boldsymbol{k}} + \chi(\boldsymbol{k}, \mathrm{i}\omega_n)) & \phi_1(\boldsymbol{k}, \mathrm{i}\omega_n) - \mathrm{i}\phi_2(\boldsymbol{k}, \mathrm{i}\omega_n) \\ \phi_1(\boldsymbol{k}, \mathrm{i}\omega_n) + \mathrm{i}\phi_2(\boldsymbol{k}, \mathrm{i}\omega_n) & \mathrm{i}\omega_n Z(\boldsymbol{k}, \mathrm{i}\omega_n) - (\xi_{\boldsymbol{k}} + \chi(\boldsymbol{k}, \mathrm{i}\omega_n)) \end{pmatrix}\end{aligned} \tag{2.21}$$

where

$$\Theta(\boldsymbol{k}, \mathrm{i}\omega_n) := (\mathrm{i}\omega_n Z(\boldsymbol{k}, \mathrm{i}\omega_n))^2 - (\xi_{\boldsymbol{k}} + \chi(\boldsymbol{k}, \mathrm{i}\omega_n))^2 - \phi_1(\boldsymbol{k}, \mathrm{i}\omega_n)^2 - \phi_2(\boldsymbol{k}, \mathrm{i}\omega_n)^2. \tag{2.22}$$

The particle and hole elementary excitations are given by the poles of the Greens function, i.e. by the condition $\det G(\boldsymbol{k}, \mathrm{i}\omega_n) = \Theta(\boldsymbol{k}, \mathrm{i}\omega_n) = 0$, leading to:

$$E_{\boldsymbol{k}} = \sqrt{\left(\frac{\xi_{\boldsymbol{k}} + \chi(\boldsymbol{k}, \mathrm{i}\omega_n)}{Z(\boldsymbol{k}, \mathrm{i}\omega_n)}\right)^2 + \frac{\phi_1(\boldsymbol{k}, \mathrm{i}\omega_n)^2 + \phi_2(\boldsymbol{k}, \mathrm{i}\omega_n)^2}{Z(\boldsymbol{k}, \mathrm{i}\omega_n)^2}}.$$

In the normal state, the off-diagonal elements of the Greens function vanish ($\phi_1(\boldsymbol{k}, \mathrm{i}\omega_n) = \phi_2(\boldsymbol{k}, \mathrm{i}\omega_n) \equiv 0$), leaving the first term under the square root as the normal-state excitation spectrum. In the superconducting state, finite $\phi_{1,2}(\boldsymbol{k}, \mathrm{i}\omega_n)$ open a gap in the excitation spectrum (analogous to BCS theory section 2.1), given by the second term; on this basis, a gap function can be written as:

$$\begin{aligned}\Delta(\boldsymbol{k}, \mathrm{i}\omega_n) :=& \frac{\phi_1(\boldsymbol{k}, \mathrm{i}\omega_n) - \mathrm{i}\phi_2(\boldsymbol{k}, \mathrm{i}\omega_n)}{Z(\boldsymbol{k}, \mathrm{i}\omega_n)} \\ |\Delta(\boldsymbol{k}, \mathrm{i}\omega_n)| =& \sqrt{\frac{\phi_1(\boldsymbol{k}, \mathrm{i}\omega_n)^2 + \phi_2(\boldsymbol{k}, \mathrm{i}\omega_n)^2}{|Z(\boldsymbol{k}, \mathrm{i}\omega_n)|^2}}.\end{aligned} \tag{2.23}$$

Both Z and χ are finite also in the normal state; while χ introduces a shift in the energy levels, Z acts as a renormalization, both to the electronic energies and to the superconducting gap.

2.2.3. Anisotropic Eliashberg equations

In Eliashberg theory, the Migdal theorem is used to approximate the phononic contribution to the electron self-energy (the Coulomb contribution will be discussed in subsection 2.2.5); its central statement is that the dominant contribution is from lowest-order diagrams (including a single phonon propagator line), and that higher-order diagrams only contribute in the order of $\omega_\mathrm{D}/E_\mathrm{F}$ (typically two orders of magnitude lower). Translated into an equation, the resulting self-energy is

$$\Sigma_\mathrm{ph}(\boldsymbol{k}, \mathrm{i}\omega_n) := -\tfrac{1}{\beta}\sum_{\boldsymbol{k}'n'} \widehat{\sigma}_3 G(\boldsymbol{k}', \mathrm{i}\omega_{n'})\widehat{\sigma}_3 \sum_\nu |g_{\boldsymbol{k}\boldsymbol{k}'\nu}|^2 D_\nu(\boldsymbol{k}-\boldsymbol{k}', \mathrm{i}\omega_n - \mathrm{i}\omega_{n'}), \tag{2.24}$$

with the electron-phonon matrix elements $g_{\boldsymbol{k}\boldsymbol{k}'\nu}$ (cf. subsection 1.2.4) and the phonon propagator

$$D_\nu(\boldsymbol{q}, \mathrm{i}\omega_n) = \frac{-2\Omega_{\boldsymbol{q}\nu}}{\omega_n^2 + \Omega_{\boldsymbol{q}\nu}^2}.$$

Expressing the interacting Greens function in (2.24) as (2.21), a comparison of coefficients on the Pauli matrices in (2.20) yields the *anisotropic Eliashberg equations*

$$[1 - Z(\boldsymbol{k}, \mathrm{i}\omega_n)]\,\mathrm{i}\omega_n = \frac{1}{\beta}\sum_{\boldsymbol{k}'n'\nu} |g_{\boldsymbol{k}\boldsymbol{k}'\nu}|^2 \mathrm{i}\omega_{n'} Z(\boldsymbol{k}', \mathrm{i}\omega_{n'}) \frac{D_\nu(\boldsymbol{k}-\boldsymbol{k}', \mathrm{i}\omega_n - \mathrm{i}\omega_{n'})}{\Theta(\boldsymbol{k}', \mathrm{i}\omega_{n'})} \tag{2.25}$$

$$\chi(\boldsymbol{k}, \mathrm{i}\omega_n) = \frac{1}{\beta}\sum_{\boldsymbol{k}'n'\nu} |g_{\boldsymbol{k}\boldsymbol{k}'\nu}|^2 \left[\chi(\boldsymbol{k}', \mathrm{i}\omega_{n'}) + \xi_{\boldsymbol{k}'}\right] \frac{D_\nu(\boldsymbol{k}-\boldsymbol{k}', \mathrm{i}\omega_n - \mathrm{i}\omega_{n'})}{\Theta(\boldsymbol{k}', \mathrm{i}\omega_{n'})} \tag{2.26}$$

$$\phi_{1,2}(\boldsymbol{k}, \mathrm{i}\omega_n) = -\frac{1}{\beta}\sum_{\boldsymbol{k}'n'\nu} |g_{\boldsymbol{k}\boldsymbol{k}'\nu}|^2 \phi_{1,2}(\boldsymbol{k}', \mathrm{i}\omega_{n'}) \frac{D_\nu(\boldsymbol{k}-\boldsymbol{k}', \mathrm{i}\omega_n - \mathrm{i}\omega_{n'})}{\Theta(\boldsymbol{k}', \mathrm{i}\omega_{n'})}, \tag{2.27}$$

where $\Theta(\boldsymbol{k}', \mathrm{i}\omega_{n'})$ was given by (2.22). The equations for ϕ_1 and ϕ_2, associated with the off-diagonal Pauli matrices, are identical and therefore, the solutions will only differ by a phase factor. However, as the equations couple all crystal momenta $\boldsymbol{k}$, the associated computational cost for a direct solution is huge.

2.2.4. One-dimensional Eliashberg equations

Computational cost can be greatly reduced by defining a new set of equations by averaging over $(\boldsymbol{k}, \boldsymbol{k}')$ within the Fermi surface $\xi_{\boldsymbol{k}} = 0$, as both Z (2.25) and $\phi_{1,2}$ (2.27) are only nonzero very close to it. The shift of the eigenvalues χ (2.26), on the other hand, is usually small and will therefore be neglected altogether from here on. An important (indirect) dependence on $\boldsymbol{k}$ is however kept: the dependence of the determinant Θ (2.22) on the eigenvalues $\xi_{\boldsymbol{k}}$. The resulting *one-dimensional Eliashberg equations* turn out to yield realistic results in the case of simple cases, but show significant discrepancies when anisotropy plays a role in the system of interest (e.g. in $\mathsf{MgB_2}$).

The averages are performed application of an operator $\frac{1}{N(0)}\sum_{\boldsymbol{k}} \delta(\xi_{\boldsymbol{k}})$, where $N(0)$ is the normal electron density of states at Fermi level, and an insertion of unity $\int \mathrm{d}\Omega \delta(\Omega -$

$\Omega_{\boldsymbol{q}\nu}$), where $\boldsymbol{q} \equiv \boldsymbol{k} - \boldsymbol{k}'$ is a phonon wave vector (conservation of crystal momentum). As a result, (2.25) and (2.27) become

$$\begin{aligned}
[1 - Z(\mathrm{i}\omega_n)]\,\mathrm{i}\omega_n = &- \frac{1}{\beta N(0)^2} \sum_{n'} \int \mathrm{d}\Omega \sum_{\boldsymbol{k}\boldsymbol{q}\nu} \frac{|g_{\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q},\nu}|^2 \delta(\xi_{\boldsymbol{k}})\delta(\xi_{\boldsymbol{k}+\boldsymbol{q}})\delta(\Omega - \Omega_{\boldsymbol{q}\nu}) 2\Omega_{\boldsymbol{q}\nu}}{(\omega_n - \omega_{n'})^2 + \Omega_{\boldsymbol{q}\nu}^2} \\
&\times \int_{-\infty}^{\infty} \mathrm{d}\xi \frac{N(\xi)\mathrm{i}\omega_{n'} Z(\mathrm{i}\omega_{n'})}{\Theta(\xi, \mathrm{i}\omega_{n'})} \\
\phi_{1,2}(\mathrm{i}\omega_n) = &\quad \frac{1}{\beta N(0)^2} \sum_{n'} \int \mathrm{d}\Omega \sum_{\boldsymbol{k}\boldsymbol{q}\nu} \frac{|g_{\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q},\nu}|^2 \delta(\xi_{\boldsymbol{k}})\delta(\xi_{\boldsymbol{k}+\boldsymbol{q}})\delta(\Omega - \Omega_{\boldsymbol{q}\nu}) 2\Omega_{\boldsymbol{q}\nu}}{(\omega_n - \omega_{n'})^2 + \Omega_{\boldsymbol{q}\nu}^2} \\
&\times \int_{-\infty}^{\infty} \mathrm{d}\xi \frac{N(\xi)\phi_{1,2}(\mathrm{i}\omega_{n'})}{\Theta(\xi, \mathrm{i}\omega_{n'})},
\end{aligned}$$

where, based on the low phonon energy scale, a separation of the $\boldsymbol{k}'$ summation into an angular average for $\xi_{\boldsymbol{k}} = 0$ and an integral over ξ on the normal state energy dependent part has been performed [74]. A further approximation can be made if $N(\xi)$ only varies slowly around E_F, substituting $N(\xi) \to N(0)$ (the integrand of the ξ-integral rapidly decays with distance from Fermi level due to the $\xi_{\boldsymbol{k}}^2$ dependence of the denominator); the integration can then be performed analytically, yielding:

$$[1 - Z(\mathrm{i}\omega_n)]\,\mathrm{i}\omega_n = -\frac{\pi}{\beta} \sum_{\omega_{n'}} \frac{\mathrm{i}\omega_{n'} Z(\mathrm{i}\omega_{n'})}{\Xi(\mathrm{i}\omega_{n'})} \int \mathrm{d}\Omega \frac{\alpha^2 F(\Omega) 2\Omega}{(\omega_n - \omega_{n'})^2 + \Omega_{\boldsymbol{q}\nu}^2} \tag{2.28}$$

$$\phi_{1,2}(\mathrm{i}\omega_n) = \frac{\pi}{\beta} \sum_{\omega_{n'}} \frac{\phi_{1,2}(\mathrm{i}\omega_{n'})}{\Xi(\mathrm{i}\omega_{n'})} \int \mathrm{d}\Omega \frac{\alpha^2 F(\Omega) 2\Omega}{(\omega_n - \omega_{n'})^2 + \Omega_{\boldsymbol{q}\nu}^2} \tag{2.29}$$

$$\Xi(\mathrm{i}\omega_n) := \sqrt{[Z(\mathrm{i}\omega_n)\omega_n]^2 + \phi_1^2(\mathrm{i}\omega_n) + \phi_2^2(\mathrm{i}\omega_n)}$$

with the *Eliashberg function*, which summarizes the total coupling for Fermi-level electrons to phonons with frequency Ω

$$\alpha^2 F(\Omega) := N(0) \sum_{\boldsymbol{q}\nu} g_{\boldsymbol{q},\nu}^2 \delta(\Omega - \Omega_{\boldsymbol{q}\nu}), \tag{2.30}$$

where

$$g_{\boldsymbol{q},\nu}^2 := \frac{1}{N(0)^2} \sum_{\boldsymbol{k}} |g_{\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q},\nu}|^2 \delta(\xi_{\boldsymbol{k}})\delta(\xi_{\boldsymbol{k}+\boldsymbol{q}}) \tag{2.31}$$

is the Fermi-surface average coupling of a single mode $\boldsymbol{q}\nu$. Inserting (2.31) in (2.30), one obtains

$$\alpha^2 F(\Omega) = \frac{1}{N(0)} \sum_{\boldsymbol{k}\boldsymbol{q}\nu} |g_{\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q},\nu}|^2 \delta(\xi_{\boldsymbol{k}})\delta(\xi_{\boldsymbol{k}+\boldsymbol{q}})\delta(\Omega - \Omega_{\boldsymbol{q}\nu}). \tag{2.32}$$

It should be further noted that the phonon-frequency integral in (2.28) and (2.29) provides a natural way to introduce a cutoff $|\omega_{n'}| < \omega_c$ in the Matsubara summation: for

one, $\alpha^2 F(\Omega) = 0$ for any Ω larger than the maximum phonon frequency in a given material; for another, the Matsubara frequency difference in the denominator gets large for any $\omega_{n'}$ far from Fermi level. Typically, $\omega_c \approx 10\omega_{\mathrm{D}}$ is sufficient for convergence.

2.2.5. Coulomb pseudopotential

What was neglected up to this point is the influence of the repulsive Coulomb interaction on the anomalous propagators. The Coulomb contribution to the electronic self-energy is defined by diagrams with only a single Coulomb propagator line

$$\Sigma_{\mathrm{C}}(\boldsymbol{k}, \mathrm{i}\omega_n) := -\tfrac{1}{\beta} \sum_{\boldsymbol{k}' n'} \widehat{\sigma}_3 G(\boldsymbol{k}', \mathrm{i}\omega_{n'}) \widehat{\sigma}_3 V_{\mathrm{C}}(\boldsymbol{k}, -\boldsymbol{k}', \omega_n - \omega_{n'}),$$

and can be readily included in the anisotropic Eliashberg equations (2.25)-(2.27). As mentioned already in that context, practical calculations require approximation into the isotropic form.

However, including the Coulomb interaction into the *isotropic* Eliashberg equations is a hard task, as the considerations leading to reduced energy integrations and Matsubara summations (by the cutoff frequency ω_c) in the phonon case are not valid for the direct Coulomb contributions. It is therefore clear that further approximations have to be taken, which means that the Coulomb repulsion cannot be treated at the same level of accuracy as the phononic contributions to the electron self-energy.

The approach taken by Morel and Anderson [72] is justified by the difference between direct Coulomb interaction (el-el) and indirect phonon-mediated attraction (el-ph) in terms of time- and energy scales: el-el has a large energy scale, while narrow time scale; el-ph has the typically much larger timescale of inverse phonon frequencies. The difference in timescales can then be used to define a renormalized Coulomb interaction having a reduced energy window, the *Morel-Anderson pseudopotential*

$$\mu^* = \frac{\mu}{1 + \mu \ln(E/\omega_c)}, \tag{2.33}$$

where $\mu = N(0) < V_{\mathrm{C}} >$, $< V_{\mathrm{C}} >$ being an average Coulomb matrix element and E is representative for the electronic energies, such as plasma frequencies [75].

As the Coulomb interaction has been already included in the normal state self-energies by the dispersion $\xi_{\boldsymbol{k}}$, which is found along the diagonal of the matrix self energy, a correction needs only to be performed on the off-diagonals $\phi_{1,2}$:

$$\phi^C_{1,2}(\mathrm{i}\omega_n) := -\mu^* \tfrac{\pi}{\beta} \sum_{\omega_{n'}} \frac{\phi_{1,2}(\mathrm{i}\omega_{n'})}{\Xi(\mathrm{i}\omega_{n'})} \theta(\omega_c - |\omega_{n'}|). \tag{2.34}$$

Using the gap function defined by (2.23), including this contribution yields

$$\Delta(\mathrm{i}\omega_n) Z(\mathrm{i}\omega_n) = \frac{\pi}{\beta} \sum_{\omega_{n'}} \frac{\Delta(\mathrm{i}\omega_{n'})}{\sqrt{\omega_{n'}^2 + \Delta(\mathrm{i}\omega_{n'})^2}} \left[\lambda(\mathrm{i}\omega_{n'} - \mathrm{i}\omega_n) - \mu^*\right] \theta(\omega_c - |\omega_{n'}|) \tag{2.35}$$

$$Z(\mathrm{i}\omega_n) = 1 + \frac{\pi}{\omega_n \beta} \sum_{\omega_{n'}} \frac{\omega_{n'}}{\sqrt{\omega_{n'}^2 + \Delta(\mathrm{i}\omega_{n'})^2}} \lambda(\mathrm{i}\omega_{n'} - \mathrm{i}\omega_n) \tag{2.36}$$

where

$$\lambda(\mathrm{i}\omega_{n'} - \mathrm{i}\omega_n) := \int_0^\infty \mathrm{d}\Omega \, \frac{2\Omega\alpha^2 F(\Omega)}{\Omega^2 + (\mathrm{i}\omega_{n'} - \mathrm{i}\omega_n)^2}$$

is a dimensionless measure of the strength of $\alpha^2 F(\Omega)$; the special case

$$\lambda := \lambda(0) = \int_0^\infty \mathrm{d}\Omega \, \frac{2\alpha^2 F(\Omega)}{\Omega} \tag{2.37}$$

is a generalization of the coupling parameter $\lambda = N_{E_\mathrm{F}} V$ (2.12) in BCS theory.

One must note, however, that the Morel-Anderson pseudopotential μ^* (2.33) is rather difficult to estimate from first principles. Consequently, it is frequently explicitely set in order to reproduce the experimentally observed critical temperature; typical values are found in the range of $0.1 - 0.16$.

2.2.6. McMillan and Allen-Dynes formulas

The solution of the isotropic Eliashberg equations (2.35-2.36) is within reach of present computational resources, however it comes at significant computational cost. The most relevant results, i.e. the critical temperature T_c, on the other hand, can be obtained at reasonable accuracy by an analytic formula derived by McMillan [76], in a similar spirit as (2.12) in the BCS case:

$$T_\mathrm{c}^\mathrm{McMillan} = \frac{\Theta_\mathrm{D}}{1.45} \exp - \frac{1.04(1+\lambda)}{\lambda - \mu^*(1 + 0.62\lambda)}, \tag{2.38}$$

where λ was defined in (2.37) and Θ_D is the Debye temperature. McMillan obtained this result by a mixed theoretical, numerical and empirical approach. A model was obtained by approximating a real-axis formulation of the Eliashberg equations [74, 77], and then fitted to data obtained from experiment [78] and computational methods for a set of superconductors known at that time.

While (2.38) was, despite its simplicity, quite successful in predicting critical temperatures at its time, Dynes [79] showed about 4 years later that many intermediately discovered materials showed large error bars; he related this problem to the choice of the Debye frequency as a representative frequency in the prefactor of the exponential. Another 3 years later Allen and Dynes proposed [75]

$$T_\mathrm{c}^\mathrm{AllenDynes} = f_1 f_2 \frac{\Omega_\mathrm{log}}{1.2} \exp - \frac{1.04(1+\lambda)}{\lambda - \mu^*(1 + 0.62\lambda)}, \tag{2.39}$$

where

$$\Omega_\mathrm{log} := \exp\left[\frac{2}{\lambda}\int_0^\infty \log(\Omega)\frac{\alpha^2 F(\Omega)}{\Omega}\,\mathrm{d}\Omega\right], \tag{2.40}$$

which takes the shape of $\alpha^2 F$ better into account than the Debye temperature Θ_D used by McMillan. Allen and Dynes showed that McMillan's equation shows wrong asymptotic behaviour in the limit of large λ (where $T_\mathrm{c} \propto \sqrt{\lambda}$ according to the Eliashberg equations),

an effect lowering prediction accuracy already at $\lambda > 1.5$; moreover, the shape of $\alpha^2F(\Omega)$ plays a larger role for larger coupling strenghts. These errors are compensated by the introduction of two correction factors f_1 and f_2, both unity for small λ:

$$f_1 := \sqrt[3]{1 + \sqrt{\lambda/(2.46 * (1 + 3.8\mu^*))}^3} \tag{2.41}$$

$$f_2 := 1 + \frac{\left(\Omega_2/\Omega_{\log} - 1\right)\lambda^2}{\lambda^2 + \left(1.82\Omega_2/\Omega_{\log}(1 + 6.3\mu^*)\right)^2} \tag{2.42}$$

introducing a second frequency parameter

$$\Omega_2 := \sqrt{\frac{2}{\lambda}\int_0^\infty \Omega\alpha^2F(\Omega)\,\mathrm{d}\Omega} \tag{2.43}$$

as an additional descriptor for the shape of the Eliashberg function $\alpha^2F(\Omega)$.

2.3. Electron-Phonon interaction and superconductivity

The central concept within our high-throughput search for superconductors are the *Descriptors of superconductivity*, to be introduced in chapter 5, which provide an estimate of superconductivity in a given material on the basis of computationally cheap normal state properties.

Therfore, we will conclude our review on the theory of superconductivity with a summary of already known relations between structural/normal state properties, the electron-phonon interaction and the critical temperature T_c.

2.3.1. Relation between critical temperature and electron-phonon coupling

As discussed in subsection 2.2.6, an analyic formula, McMillan's equation (2.38), can be determined by careful fitting, which predicts the critical temperature from isotropic Elishberg theory. The more accurate Allen-Dynes formula (2.39) based on the former shows an RMS error of about 5% on the set of 300 materials accessible to authors [75], and by the means of asymptotic correction terms (2.41, 2.42) also recovers the theoretical asymptotic limit $T_c \propto \sqrt{\lambda}$ for large λ.

The analytic formula only depends on a set of 4 parameters: λ, $\Omega_{\log}$, $\bar{\Omega}_2$ and μ^*. λ (2.37) describes the total strength of the electron-phonon coupling, $\Omega_{\log}$ (2.40) and $\bar{\Omega}_2$ (2.43) are representative phonon frequencies, the latter used to account for the shape of the Eliashberg function $\alpha^2F(\Omega)$ and only contributing in the asymptotic correction factor; as $\alpha^2F(\Omega)$ is accessible from first principles (section 1.2), the first three parameters are so as well. Direct Coulomb repulsion is accounted for by the Morel-Anderson pseudopotential μ^*, which for conventional superconductors lies within the range between 0.1 and 0.16; this parameter is hard to evaluate ab-initio, and is conventionally used as a fitting parameter to match the experimental T_c.

Before switching to a deeper investigation of λ, let us investigate the λ and μ^* dependence of the McMillan-T_c, on the example of Nb_3Sn, illustrated in Fig. 2.1: for a constant

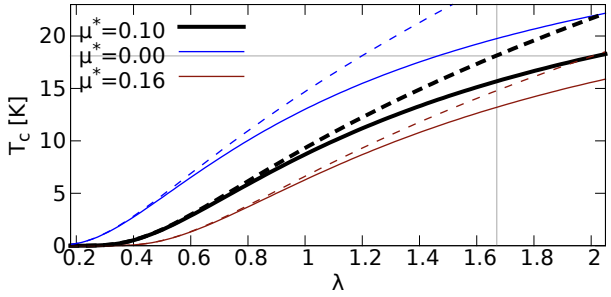


Figure 2.1.: Superconducting $T_c(\lambda)$ according to the Allen-Dynes variant (2.39) of McMillan's equation (phonon frequency parameters from Nb_3Sn). Dashed curves include the Allen-Dynes asymptotic correction terms. Coulomb pseudopotential $\mu^* = 0.10$ is realistic [75].

μ^*, T_c is monotonically increasing with the electron-phonon coupling parameter λ (the realistic one for Nb_3Sn is marked by a gray line); it can be observed that the asymptotic correction term $f_1 f_2$ has neglegible effect in the small-λ regime (dashed curves).

The fact of T_c monotonically increasing with λ is independent of the actual values of μ^* and the phonon frequency parameters $\Omega_{\log}$, $\bar{\Omega}_2$ (where the black curves for $\mu^* = 0.1$ correspond to the Nb_3Sn data from [75]).

Therefore it is a central objective in our high-throughput search for conventional superconductors (Part II) to detect materials with λ sufficiently large to support superconductivity at significant transition temperatures. Anticipating chapter 5: while λ can be evaluated numerically via density functional perturbation theory (section 1.2), such calculations are too computationally expensive to be considered in a high-throughput context. A central goal of the present work is it therefore to find estimates for λ just from structural and electronic ground-state properties.

In preparation for the later discussion, we will outline already known properties of the electron-phonon coupling strength λ within this section.

2.3.2. Properties of the electron-phonon coupling strength

Relation to the Density of States at Fermi Level

As a first step, let us assume an Einstein model for the phonons, which describes the ions as an ensemble of independent harmonic oscillators, which leads to a flat phonon dispersion $\omega^{\mathrm{E}}_{\boldsymbol{q},\nu} := \omega^{\mathrm{E}}$. Furthermore, let us assume $g^{n,n'}_{\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q},\nu} = g$, i.e. neither anisotopy nor nesting effects are present. Combining these assumptions with (2.32), (2.37), we obtain

$$\lambda^{\mathrm{E}} = \frac{1}{\omega^{\mathrm{E}}} \frac{1}{N(0)} g^2 N(0)^2 = \frac{g^2}{\omega^{\mathrm{E}}} N(0), \tag{2.44}$$

which means that the electron-phonon coupling constant λ is proportional to the electronic density of states at Fermi level ($\mathrm{DOS}_{\mathrm{F}} := N(0)$), in agreement with the BCS approach (2.12).

Deformation potential

In our Einstein model, the expression (2.44) can be further split into subparts concerning dynamic properties, such as ionic mass and frequency; note that g^2, definition (1.31) both explicitely depends on the frequency, and implicitely, through the definition of the normal mode displacement patterns in a general polyatomic system: in this formulation, displacement of each ion I is scaled by $\sqrt{M_I}^{-1}$, where M_I is the nuclear mass. Assuming all nuclear masses in our system were identical $M_I = M$, this scaling can be included within the amplitude prefactor, such that

$$\tilde{g}^{\nu,n,m}_{\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q}} = \underbrace{\sqrt{\frac{\hbar}{2\omega_{\boldsymbol{q}\,\nu} M}}}_{\text{amplitude factor}} \underbrace{\left\langle \varphi_{\boldsymbol{k}+\boldsymbol{q}\,m} \middle| \Delta_{\boldsymbol{q}\,\nu} v_s \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \middle| \varphi_{\boldsymbol{k}\,n} \right\rangle}_{\text{deformation potential term}}. \tag{2.45}$$

We call the last term *deformation potential term*, as it is related to the phonon deformation potential [80, 81], a tensor describing the electronic perturbation due to *unit* lattice deformation, while the magnitude of the perturbation is approximated to scale linearly with the amplitude of ion displacement.

Incorporating this consideration,(2.44) can be written as

$$\lambda^{\mathrm{E}} = \tfrac{1}{M\omega^2} N(0) I^2 \tag{2.46}$$

where I is the Fermi-surface average deformation potential term for all phonon modes $\boldsymbol{q}\nu$. It is a central part of the present work to establish a relation between the electronic structure of the system in equilibrium lattice configuration and the magnitude of I^2 (section 5.3).

3. Machine Learning

As part of the present work, we have developed and applied machine learning methods to directly predict electronic properties just from the crystal structure. Kohn-Sham DFT calculations (section 1.1) could then be substituted by these computationally far cheaper predictions within a high-throughput search for superconductors. Although the machine learning aproach is not a central part of this work, we present a review of the methods applied; despite their growing importance also among physicists, such methods cannot be considered well-known.

Machine learning is a branch of the scientific field of artificial intelligence. Its topic is to find algorithms that, given a set of data,

- characterize the data quantitatively, finding the features of the underlying mechanisms that lead to the data
- use this characterization to predict properties of unknown data

Machine learning (ML) methods are widely and successfully used in practical applications. For example, in optical character recognition (OCR) applications, used to sort mail by possibly hand-written postal codes; the best algorithms in the field have an error rate of single-digit recognition comparable to the human error rate (about 2.5%) [82]. ML as such is universal, and the same methods are successfully applied to vastly different problems, for example: the aforementioned OCR, automatic image classification, speech recognition, brain-computer-interfaces [83]

All algorithms used within the present work belong to the class of *supervised learning methods*: *predictors* for Kohn-Sham-DFT electronic properties are *trained* on a set of materials where these properties have been explicitely computed. The predictors can consecutively be used to access electronic properties of any new material *without* having to perform a KS-DFT calculation.

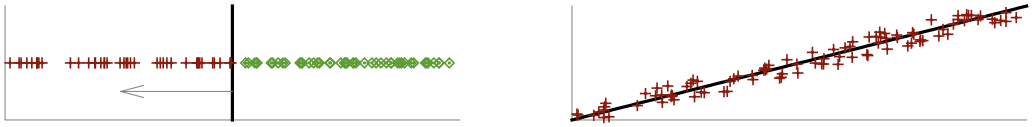
Supervised learning is performed on a *training set*

$$\mathcal{D} = \{(\boldsymbol{t}_j, d_j)\} \mid \boldsymbol{t}_j \in \mathcal{X}, d_j \in \mathcal{Y}$$

with *input space* $\mathcal{X}$ and *label space* $\mathcal{Y}$. The *known* labels d_j of the training set are called the *supervisory signal*. It is the goal of a supervised learning algorithm to find a *predictor*

$$f_{\mathcal{D}} : \mathcal{X} \to \mathcal{Y},$$

based on the training data $\mathcal{D}$, which predicts labels also for input not part of the original training set.



(a) Classification: label by color **(b)** Regression: label on y axis

Figure 3.1.: Linear classification and regression

3.1. Input Space $\mathcal{X}$

The input space $\mathcal{X}$ is required to be a Hilbert space by the methods presented in this chapter. Finding a good *representation* of input objects, such as pictures, sounds, molecules or in our case crystal structures as members of $\mathcal{X}$ is the first and crucial step in applying machine learning methods.

3.2. Linear predictors

While *linear* models may seem too simple for practical use, a powerful method, presented in section 3.3, exists which maps nonlinear models to linear ones, making them actually a powerful tool in machine learning.

Depending on the label space $\mathcal{Y}$, different names are used for the predictors, and also the training algorithms differ strongly:

Classifier is the term used in the context of discrete $\mathcal{Y}$. We restrict our discussion to *binary classification* with labels either $+1$ or -1. In such a case, a linear classifier is characterized by

$$f(\boldsymbol{x}) = \operatorname{sgn}(\boldsymbol{w} \cdot \boldsymbol{x} + b), \tag{3.1}$$

interpreted as a hyperplane in $\mathcal{X}$ separating the two classes (Figure 3.1a). Such a classifier is also called a *perceptron*, it was introduced as a simplified model of a neuron [84].

Regression function is the term used in the context of continous $\mathcal{Y}$, which corresponds to a straight line in $\mathcal{X} \otimes \mathcal{Y}$ (Figure 3.1b):

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b. \tag{3.2}$$

3.2.1. Predictor Training

Model parameters $(\boldsymbol{w}, b)$ are determined by *predictor training* in both cases, specific algorithms are presented in section 3.4. Such algorithms are, from a mathematical point of view, constrained optimization problems, which provide a compromise between global prediction error $E_{\mathcal{D}}(f_{\mathcal{D}})$ on the training set $\mathcal{D}$ and the *complexity* of the model[82, 85, 86]: in case of high $\dim(\boldsymbol{w})$, which is common when applying the nonlinear extensions

(section 3.3), the large number of parameters may lead to *overfitting*. The term "overfitting" refers to models which minimize the empirical risk, i.e. minimize the prediction error on $\mathcal{D}$, but fail to generalize, i.e. have a high upper bound of prediction error when applied to unknown data (generalization error, discussed in the next subsection). Within the training algorithms, a *regularization term* is included in the optimization process, which penalizes more complex models, improving the predictor's generalization properties.

3.2.2. Generalization error and cross validation

As emphasized before, an important feature of a *good* predictor is its capability to *generalize*, i.e. its ability to correctly predict the labels d'_j associated to input $\boldsymbol{x}_j$ *not seen* during the training phase; therefore the quality of a predictor $f_{\mathcal{D}}(\boldsymbol{x})$, trained on set $\mathcal{D}$ can be estimated from the prediction error within a *test set* $\mathcal{T} = \{(\boldsymbol{x_j}, d'_j)\},\ j = 1 \dots M$, again with known labels d'_j, but disjoint from the training set $\mathcal{D}$:

$$E_{\mathcal{T}}^{\mathrm{MAE}}(f_{\mathcal{D}}) = \frac{1}{M} \sum_{j=1}^{M} \left| f_{\mathcal{D}}(\boldsymbol{x}_j) - d'_j \right| .$$

Cross validation provides a method to test for the applicability of a prediction model, based on the generalization error. In a straightforward implementation, a set $\mathcal{D}'$ of data with known labels is split into L subsets $\mathcal{D}'_j$ of equal size. Training of L individual predictors is performed on training sets $\mathcal{D}_i = \cup_{j \neq i} \mathcal{D}'_j$, while generalization errors are evaluated within $\mathcal{T}_i = \mathcal{D}'_i$. The special case of $L = |\mathcal{D}'|$, i.e. training on all-but-one members of $\mathcal{D}'$, while testing on one member, is referred to as *Leave-one-out cross-validation* (LOOCV).

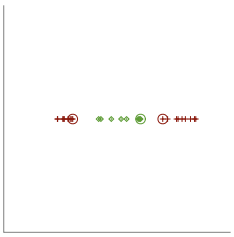
An average of the subset errors $E_{\mathcal{T}_i}(f_{\mathcal{D}_i})$ provides information on the quality of the underlying statistical model itself, as it is independent of the concrete training set of a single predictor. Due to this property, cross-validation provides a path for the optimization of the statistical model itself (as opposed to the model training subsection 3.2.1).

3.3. Feature space and Kernel trick

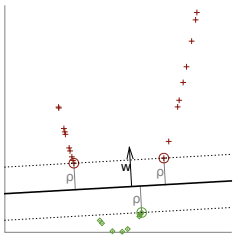
The classifier/perceptron (3.1) and the regression function (3.2) can only be applied to linear problems; in the case of the perceptron, this means that the two classes need to be separable by a hyperplane in input space $\mathcal{X}$. Throughout this section, we use the classification problem as an illustrative example, while the same argumentation holds for the regression problem.

The approach described within this section extends the applicability of linear predictors to nonlinear problems, by mapping the input data is to a potentially much-higher dimensional *feature space* $\mathcal{F}$ via a nonlinear transformation

$$\boldsymbol{\phi}: \quad \mathcal{X} \to \mathcal{F}.$$

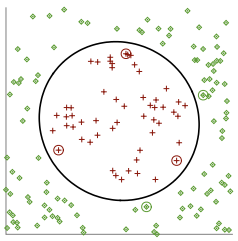


(a) Input space

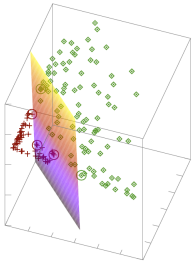


(b) Feature space

Figure 3.2.: $\phi(\boldsymbol{x}) = (x_1, x_1^2)$



(a) Input space



(b) Feature space

Figure 3.3.: $\phi(\boldsymbol{x}) = (x_1, \sqrt{2}x_1x_2, x_2^2)$

	$\boldsymbol{\phi}(\boldsymbol{x})$	$\dim \mathcal{F}$	$k(\boldsymbol{x}, \boldsymbol{y})$
Polynomial, degree d	$\left\{\prod_{i=1}^{n} x_i^{b_i} \mid \sum_{i=1}^{n} b_i \leq d\right\}$	$\binom{n+d-1}{n}$	$(\boldsymbol{x} \cdot \boldsymbol{y} + \theta)^d$
Radial basis functions			
Gaussian, width σ	$\left\{\mathrm{e}^{-\frac{1}{2\sigma^2}\lvert\boldsymbol{x}-\boldsymbol{c}\rvert^2} \mid \forall \boldsymbol{c} \in \mathcal{X}\right\}$	∞	$\mathrm{e}^{-\frac{1}{2\sigma^2}\lvert\boldsymbol{x}-\boldsymbol{y}\rvert^2}$
Laplacian, width σ	$\left\{\mathrm{e}^{-\frac{1}{\sigma}\lvert\boldsymbol{x}-\boldsymbol{c}\rvert} \mid \forall \boldsymbol{c} \in \mathcal{X}\right\}$	∞	$\mathrm{e}^{-\frac{1}{\sigma}\lvert\boldsymbol{x}-\boldsymbol{y}\rvert}$

Table 3.1.: Examples of feature maps and associated kernel functions

Linear separation is then performed in $\mathcal{F}$, transforming (3.1) into

$$f(\boldsymbol{x}) = \operatorname{sgn}\left(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b\right) \quad |\boldsymbol{w} \in \mathcal{F}. \tag{3.3}$$

Figure 3.2a presents an example of a data not linearly separable in the 1-dimensinonal input space $\mathcal{X}$. However, by mapping to a feature space via $\boldsymbol{\phi}(\boldsymbol{x}) = (x_1, x_1^2)$, the data becomes linearly separable (Figure 3.2b). The task discussed in the following is how to setup the mapping $\boldsymbol{\phi}$.

The approach consists in trying to generate a big number of features using a mapping function, and assuming that the perceptron is able to perform the linear separation in this high(er)-dimensional feature space. A few examples of such transformations are presented in the 2nd column of Table 3.1. Dimensionality of the radial basis function (RBF) feature spaces is infinite, as they correspond to projections on RBF centered on each possible point in $\mathcal{X}$; consequently, their only practical use requires the *kernel trick* presented in the next subsection. The polynomial feature map suffers from similar difficulties when applied to real-world problems: in the case of optical character recognition (OCR) on grayscale images with $X = 28 \times 28$ pixels, polynomial features with degree $d = 7$ are needed for good performance [87]. This feature space has about $3.7 \cdot 10^{16}$ features, making also this polynomial feature space infeasable to use without the kernel trick.

3.3.1. The Kernel Trick

Feature extraction can be expensive, and the dimensionality of feature space can be huge, which again raises the costs for practical calculations. In the case of the RBF feature map Table 3.1, the feature space itself is infinite-dimensional, therefore it would be impossible to use it directly. The *Kernel Trick* is a method to reduce these costs (without giving up the benefits of the mapping to feature space).

A *kernel function* $k(\boldsymbol{x}, \boldsymbol{x}') : \mathcal{X} \otimes \mathcal{X} \to \mathbb{R}$ is a function, symmetric in its arguments, if for a feature map $\boldsymbol{\phi}$ [88, 89]:

$$k_\phi(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x}') \cdot \boldsymbol{\phi}(\boldsymbol{x}). \tag{3.4}$$

If $k(\boldsymbol{x}, \boldsymbol{x}')$ is computationally cheaper (in terms of memory and time) than the explicit computation of $\boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}')$, it can be used to calculate inner products between vectors $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, implicitely mapped to $\mathcal{F}$. Given such a kernel function, *every* linear algorithm written in terms of inner products in $\mathcal{X}$ can be extended to a non-linear variant by replacing every inner product by a kernel function.

Kernels corresponding to the previously mentioned commonly used feature maps are found in the right column of Table 3.1. As can be seen from the table, the computational cost is drastically reduced in all three cases: inner products within a Gaussian or Laplacian feature map reduce to a single evaluation of a Gaussian or Laplacian on the cartesian distance between the two points in $\mathcal{X}$, and also the previously mentioned example of polynomial features in OCR reduces essentially to an inner product in $\mathcal{X}$, avoiding an explicit evaluation of $3.7 \cdot 10^{16}$ features.

Besides these commonly used kernels, there exists a variety of other kernels, as listed in [90]. Many of them are tailored to a specific problem (DNA analysis, OCR, image classification, etc.).

3.3.2. Conclusion

Problems, which are not linear in input space $\mathcal{X}$ become linear by mapping to an appropriately choosen feature space $\mathcal{F}$.

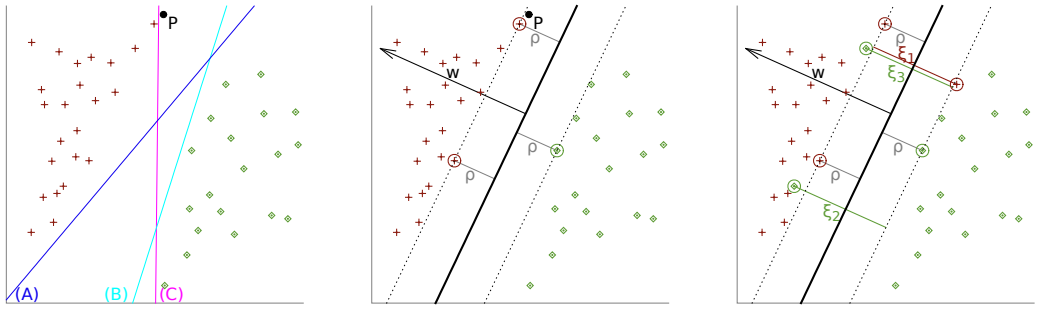
Every linear algorithm written in terms of inner products in $\mathcal{X}$ can be efficiently extended to a non-linear one by using a kernel function $\mathrm{k}(\boldsymbol{x}, \boldsymbol{x}')$, which implicitely evaluates inner products in a feature space $\mathcal{F}$ without calculating (or even knowing an explicite) feature map ϕ.

Many different kernels (and therefore feature maps) have been published. The choice of which one can be used to solve a particular problem depends on the properties of its input vectors (such as translation and rotation invariance).

3.4. Learning Algorithms

3.4.1. Classification: Support Vector Machine (SVM)

In classification problems (3.1), the choice of a separating hyperplane with normal vector $\boldsymbol{w}$ and threshold b is not unique; example planes A, B and C in Figure 3.4a are valid



(a) Linear classifiers **(b)** SVM classifier **(c)** SVM *soft-margin* classifier

Figure 3.4.: Linear classifiers: Support Vector machine. Training data is represented by red and green points, color indicates class label.

separators of the training data. However, *generalization properties* do strongly differ: classifier C would assign test point $\boldsymbol{P}$ to the 'green' class, a false prediction given the distribution of the training data.

The *support vector machine* (SVM) is a supervised learning algorithm with *optimal* (proof in [85]) generalization properties, which is the hyperplane providing the *maximum margin* ρ to either class of training data (Figure 3.4b). In the case of the SVM, exactly this condition is the *regularization term*, with the central purpose of minimizing the model complexity [82].

Normalization of the plane equation is performed with respect to the training examples $\boldsymbol{t}_j^{\pm 1,\mathrm{s}}$ found, symmetrically, in both classes closest to the hyperplane:

$$(\pm 1)(\boldsymbol{\phi}(\boldsymbol{t}_j^{\pm 1,\mathrm{s}}) \cdot \boldsymbol{w} + b) \equiv 1 \tag{3.5}$$

$$\text{defining the } \textit{margin} \text{ as } \rho = \frac{1}{2}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \cdot (\boldsymbol{\phi}(\boldsymbol{t}_j^{+1,\mathrm{s}}) - \boldsymbol{\phi}(\boldsymbol{t}_k^{-1,\mathrm{s}})) = \frac{1}{\|\boldsymbol{w}\|}$$

maximizing the margin therefore corresponds to a constrained quadratic optimization

$$\min_{\boldsymbol{w},b} \boldsymbol{w}^2 \quad \text{subject to} \quad d_j(\boldsymbol{\phi}(\boldsymbol{t}_j) \cdot \boldsymbol{w} + b) \geq 1 \; \forall (\boldsymbol{t}_j, d_j) \in \mathcal{D} \tag{3.6}$$

solved via Karush-Kuhn-Tucker (KKT) method, which extends Lagrangian multipliers to *inequality* boundary conditions. The Lagrangian of the problem reads

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \tfrac{1}{2}\boldsymbol{w} \cdot \boldsymbol{w} - \sum_{l=1}^{n} \alpha_l (d_l(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{t}_l) + b) - 1) \tag{3.7}$$

and is to be minimized with respect to $\boldsymbol{w}$ and b, and maximized with repect to the Lagrangian multipliers $\alpha \geq 0$. The minimization leads to the saddle point equations

$$0 = \frac{\partial L}{\partial b} = \sum_{l=1}^{n} \alpha_l d_l \text{ and } 0 = \boldsymbol{\nabla}_{\boldsymbol{w}} L = \boldsymbol{w} - \sum_{l=1}^{n} \alpha_l d_l \boldsymbol{\phi}(\boldsymbol{t}_l)$$

$$\Longrightarrow \quad \boldsymbol{w} = \sum_{l=1}^{n} \alpha_l d_l \boldsymbol{\phi}(\boldsymbol{t}_l). \tag{3.8}$$

showing that $\boldsymbol{w}$ is a linear combination of the training input vectors $\boldsymbol{t}_l$, mapped to $\mathcal{F}$. Substituting this expansion into (3.7) yields the dual quadratic optimization problem

$$\max_{\boldsymbol{\alpha}} \sum_{l=1}^{n} \alpha_l - \frac{1}{2} \sum_{k=1,l=1}^{n,n} \alpha_k \alpha_l d_k d_l k(\boldsymbol{t}_k, \boldsymbol{t}_l) \text{ subject to } \alpha_i \geq 0 \text{ and } \sum_i \alpha_i d_i = 0. \quad (3.9)$$

The coefficients α_i resulting from the optimization process determine the hyperplane normal $\boldsymbol{w}$ by (3.8); substitution of this expression in (3.1) defines the SVM predictor

$$f(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_{i=1}^{n} d_i \alpha_i k(\boldsymbol{t}_i, \boldsymbol{x}) + b\right) \quad (3.10)$$

In most cases, there will be only a small number N of training examples exactly on the margin (3.5). As illustrated in [82], such examples are actually the only ones with non-vanishing coefficients α_i; they are referred to as *support vectors*. Due to $N \ll |\mathcal{D}|$, the optimization problem is sparse.
The threshold b can be determined with the help of the support vectors, improving numerical stability by evaluating the set average:

$$b = \frac{1}{N} \sum_{j=1}^{N} \left(\frac{1}{d_j^{\mathrm{s}}} - \sum_{i=1}^{N} d_i^{\mathrm{s}} \alpha_i^{\mathrm{s}} k(\boldsymbol{t}_i^{\mathrm{s}}, \boldsymbol{t}_j^{\mathrm{s}})\right).$$

Soft margins are an important extension of SVM concerning *noisy* data, i.e. training data containing a low number of outliers (such as the points/distance vectors marked by ξ_i in Figure 3.4c). This is done by a relaxation of the boundary condition in (3.6) with the help of *slack variables* ξ_i, which quantify the training error; the training error $\sum_i \xi$ is then included minimization process:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\boldsymbol{w} \cdot \boldsymbol{w} + C \sum_{i=1}^{n} \xi_i \text{ subject to } d_j(\boldsymbol{t}_j \cdot \boldsymbol{w} + b) \geq 1 - \xi_i \; \forall (\boldsymbol{t}_j, d_j) \in \mathcal{D}, \; \xi_i \geq 0, \quad (3.11)$$

where C the parametrizes the compromize between margin maximalization and training error. This choice for the slack penalty introduces an upper bound on the Lagrange multipliers α_i, but leaves the dual problem (3.9) otherwise unchanged:

$$\max_{\boldsymbol{\alpha}} \sum_{l=1}^{n} \alpha_l - \tfrac{1}{2} \sum_{k=1,l=1}^{n,n} \alpha_k \alpha_l d_k d_l \langle \boldsymbol{x}_k | \boldsymbol{x}_l \rangle \text{ subject to } 0 \leq \alpha_i \leq C, \; \sum_i \alpha_i d_i = 0.$$

3.4.2. Kernel Ridge regression (KRR)

Kernel Ridge Regression [91] is an algorithm to implement a *regression function*, i.e. a predictor for *continuous* labels. For a previously unseen argument tuple, the algorithm predicts the function value.

Least-Squares regression has been developed independenty by Gauss and Legendre around the year 1800. The conventional derivation of the least-squares regression starts from the choice of a model, e.g.

$$f(\boldsymbol{x}) = w_1 \cdot x_1^2 + w_2 \cdot x_1 + w_0 \cdot 1.$$

With the perspective of the previous sections, especially the introduction of feature space in mind, one can associate the selection of a model in the conventional linear regression with the introduction of a finite-dimensional feature map $\phi : \boldsymbol{x} \to \boldsymbol{\phi}(\boldsymbol{x})$:

$$f(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{w}. \tag{3.12}$$

In order to determine the coefficients $\boldsymbol{w}$ of the linear model, the sum of squared residuals

$$S(\boldsymbol{w}) = \sum_j^n r_j^2 = \sum_j^n (d_j - \sum_k \phi_{j,k} w_k)^2 \quad = \left\| \boldsymbol{d} - \widehat{\phi}\boldsymbol{w} \right\|^2$$

with $\phi_{j,k} := \phi_k(\boldsymbol{t}_j)$ is minimized:

$$\min_{\boldsymbol{w}} S(\boldsymbol{w}) \Rightarrow \nabla_{\boldsymbol{w}} S(\boldsymbol{w}) \equiv \boldsymbol{0} \Rightarrow \boldsymbol{w} = \widehat{\phi}^{-1}\boldsymbol{d},$$

solvable if $\widehat{\phi}$ is invertible and well conditioned.

Kernel ridge regression There are two major drawbacks with the conventional linear regression method: the inverse $\widehat{\phi}^{-1}$ may be numerically unstable (due to an ill-conditioned problem), and if the feature space $\mathcal{F}$ is sufficiently complex, overfitting occurs.

Kernel Ridge Regression (KRR) addresses both of these problems by including *Tikhonov regularization* (regularization parameter λ) in the minimization, which penalizes the norm of the weight vector, leading to the constrained minimization problem:

$$\min_{\boldsymbol{w},\boldsymbol{r}} \frac{1}{2}\boldsymbol{r} \cdot \boldsymbol{r} + \frac{1}{2}\lambda \boldsymbol{w} \cdot \boldsymbol{w} \text{ subject to } r_j = \boldsymbol{\phi}(\boldsymbol{t}_j) \cdot \boldsymbol{w} - d_j.$$

The corresponding Lagrangian with the Lagrange multipliers α_i reads:

$$\mathcal{L} = \frac{1}{2}\boldsymbol{r} \cdot \boldsymbol{r} + \frac{1}{2}\lambda \boldsymbol{w} \cdot \boldsymbol{w} + \sum_j \alpha_j \left(\boldsymbol{\phi}(\boldsymbol{t}_j) \cdot \boldsymbol{w} - d_j - r_j \right) \tag{3.13}$$

$$\text{optimization: } \max_{\boldsymbol{\alpha}} \min_{\boldsymbol{w},\boldsymbol{r}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{r}, \boldsymbol{\alpha})$$

We will first perform the inner minimization:

$$\nabla_{\boldsymbol{w}} \mathcal{L} = \lambda \boldsymbol{w} + \sum_j \alpha_j \boldsymbol{\phi}(\boldsymbol{t}_j) \equiv 0 \quad \frac{\partial \mathcal{L}}{\partial r_j} = r_j - \alpha_j \equiv 0$$

$$\Rightarrow r_j = \alpha_j, \quad \boldsymbol{w} = -\frac{1}{\lambda} \sum_j \alpha_j \boldsymbol{\phi}(\boldsymbol{t}_j) \tag{3.14}$$

The weight vector $\boldsymbol{w}$ is a linear combination of the training inputs in feature space. Substituting this expression for $\boldsymbol{w}$ back into (3.12) yields

$$f(\boldsymbol{x}) = -\frac{1}{\lambda} \sum_j \alpha_j \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{t}_j) = -\frac{1}{\lambda} \sum_j \alpha_j k(\boldsymbol{x}, \boldsymbol{t}_j).$$

Substituting (3.14) back into (3.13) allows us to state the optimization problem for $\boldsymbol{\alpha}$:

$$\begin{aligned}\max_{\boldsymbol{\alpha}} & -\frac{1}{2}\boldsymbol{\alpha}\cdot\boldsymbol{\alpha} - \frac{1}{2\lambda}\sum_{i,j}\alpha_i\alpha_j k(\boldsymbol{t}_i,\boldsymbol{t}_j) - \boldsymbol{\alpha}\cdot\boldsymbol{d} \\ =\max_{\boldsymbol{\alpha}} & -\frac{1}{2}\langle\boldsymbol{\alpha}|\boldsymbol{\alpha}\rangle - \frac{1}{2\lambda}\boldsymbol{\alpha}^\top\widehat{K}\boldsymbol{\alpha} - \langle\boldsymbol{\alpha}|\boldsymbol{d}\rangle\end{aligned}$$

with the *kernel matrix*

$$\widehat{K}:\ K_{i,j} = k(\boldsymbol{t}_i,\boldsymbol{t}_j).$$

3.5. Coulomb Matrix representation of molecules

The machine-learning methods presented in this chapter rely on a representation of input data as vectors, which are members of a Hilbert space, commonly called *input space*. Therefore, for each problem to be treated by such methods, such a representation needs to be defined. In this section, we present a brief review of a representation for molecules, a problem that is related to the present work, as both molecules and crystal structures can be interpreted as subspaces of an (abstract) chemical compound space $\mathcal{C}$. Therefore, insight gathered from this molecular representation will be applied when defining our representation of crystal structures in section 6.3.

A molecule is fully characterized by the set of cartesian coordinates and nuclear charges of all constituent atoms $\{(\boldsymbol{R},Z)_I\}$. However, given that the property to be predicted is invariant under rotations and translations, there are infinitely many possible representations of any molecule, as cartesian components do *not* share such invariances. The approach of *directly* employing such a set as input vectors to an RBF kernel is of questionable use, as it would imply a minimization over all possible translations, rotations and permutations of atom indices.

The *Coulomb matrix representation* [92, 93] $\mathrm{C}^{\mathrm{mol}}$ recovers rotational and translational invariance, as its elements are defined as

$$c^{\mathrm{mol}}_{IJ} = \begin{cases} 0.5Z_I^{2.4} & \text{for } I \equiv J \\ \frac{Z_I Z_J}{|\boldsymbol{R}_I - \boldsymbol{R}_J|} & \text{for } I \neq J \end{cases}, \tag{3.15}$$

corresponding to point-charge Coulomb potential energies in the off-diagonal, providing information about geometry, while the diagonal provides information about the individual nuclear charges.

The representation was tested [92, 93] on hydrogen-saturated organic molecules, with up to 7 heavier atoms from the set { C, N, O, S }, and found to perform well in the prediction of atomization energies.

Trivial algebra shows that the orginal $\boldsymbol{r}_I$ can be reconstructed from the Coulomb matrix, up to a global rotation or translation.

Permutations of atoms

However, a fundamental symmetry is not covered directly within this representation: the individual matrix elements' numerical values depend on the *arbitrary* choice of atom

index order, reflected in the order of rows and columns in CM, while, trivially, all physical properties are invariant under any permutation of atom indices.

Therefore, given a possible representation $\mathrm{C}_i^{\mathrm{mol}}$ of a molecule $\boldsymbol{m}_i$, also the whole set

$$\mathcal{M}_i = \widehat{P}_n \mathrm{C}_i^{\mathrm{mol}} \mid n = 0 \ldots n_{\max}! \tag{3.16}$$

where $\widehat{P}_n$ are atom index permutation operators, acting simultaneously on rows and columns in CM representation, describes the same molecule. In principle, this finding can be accounted for by either including every member of $\mathcal{M}_i$, with the same associated label in the training set, or by implicitely choosing the permutiation resulting in the shortest Euclidian distance while evaluating the RBF distance measure; in practise, however, the fact that each training molecule would imply $n_{\max}!$ explicite or implicite training samples, negates most advantage (regarding computational complexity) of the ML approach over a KS-DFT calculation.

Refs [92, 93] discuss strategies to recover applicability of the representation. For one, the representation itself could be replaced by sorted eigenspectra $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots \lambda_n)$ with $\lambda_i \geq \lambda_{i+1}$ and $\mathrm{C}^{\mathrm{mol}} \boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$, at the expense of information loss. For the other, samples of $\mathcal{M}_i$ could be included in the training set and be employed when evaluating the distance measure.

3.6. Summary

In this chapter, we reviewed *kernel methods* in the field of *machine learning* (ML). The *linear predictors* are conceptually simple, but gain enormous power from the introduction of *feature space*. *Kernel functions* enable to use this power at only moderately elevated computational cost compared to the linear algorithms. At the same time, *regularization terms* within the learning process provide an effective and controllable way to prevent *overfitting*, and therefore greatly improve the qualitity of predictions on new data, which were not part of the training process.

The crucial point when applying ML methods within a new field of endavour consists in finding an *input representation* for the objects predictions will be made for.

We propose a representation for periodic solids in chapter 6, and evaluate its performance in chapter 9.

This chapter closes the review part of this work. In the next part, we present methods developed for our high-throughput search for superconductors.

Part II.

High-Throughput search for superconductors

4. Introduction

Today, high-throughput methods (HTM) are most well-known through their application in pharmaceutical research [22–24], which will serve as an example to explain the concept. In pharmaceutical HTM, huge libraries of substances are tested for desired properties, which in a simple case can be the effect on a protein related to a certain disease. Each test consists of the application of one substance (or one combination of substances) to one sample of the protein and a quantitative measurement of the effects. In the HTM approach, robots are used to perform such experiments simultanously and at a very high frequency (hence the name *high-throughput*). This process leads to the generation of vast databases. These are then investigated using data mining techniques in order to extract knowledge from the data, which in our example could mean the combination of substances having the largest, desired effect. The term **Data mining** refers to data analysis techniques in the fields of artificial intelligence, machine learning (chapter 3), statistics and database systems [94], However, experimental HTM is limited by the fact that an automatic synthesis of new substances is today only possible in a limited number of fields [95].

Computational HTM follow the same basic idea as their experimental counterpart, but avoid the limits set by the experimental synthesis of new substances. Independent ab-initio calculations of physical properties are performed on a huge number of materials.All tasks must be handled automatically [96, 97], which can be seen as an analog of the robot used in experimental HTM. As in the experimental case, properties are then collected in databases [98, 99] and investigated with data mining methods.

As calculations are performed on huge numbers of materials, HTM have to rely on computationally cheap individual calculations, restricted by the computer resources available. One of the common approaches is the introduction of *descriptors* [100, 101], quantities available at low computational cost but providing an estimate on the computationally expensive original problem. Furthermore, the desired low computational cost usually implies a reduction of accuracy, which needs to be nevertheless high enough to allow for a statistical evaluation. The latter is used to select candidates for in-depth studies, by means of more accurate computational methods and/or experimental research.

Therefore, our approach to a computational high-throughput search for superconductors requires, as a first task the determination of *descriptors of superconductivity* (chapter 5), that, while accessible at low computational cost, convey enough information to select candidate materials for in-depth research. Our approach to develop such descriptors is based both on the theoretical knowledge presented in chapter 2 and on empirical data, while connections between the former and the latter are demonstrated by

models. Tests on the descriptors applied during a small-scale high throughput screening[1] are made (chapter 8, chapter 10) and based on their predictions, a number of superconductors are identified.

Machine learning methods are evaluated as an alternative to the explicit calculation of the descriptors in chapter 6, chapter 9, with the help of the data generated in our high-throughput search.

In our search, materials are taken from a crystal structure database, the *Inorganic Crystal Structure Database* (ICSD). The methods and results of a statistical analysis on ICSD are reported in chapter 7, characterizing the library of materials available to our high-throughput search, providing information for crystal structure prediction and proposing a new chemical scale.

[1] A summary on the computational implementation of out HTM is given in Appendix A, which spans both numerical considerations for a computational evaluation of our descriptors of superconductivity, and more technical ones, such as the automatic setup and the supervision of the simulations.

5. Descriptors of Superconductivity

The critical temperature T_c of a phonon-driven superconductor can be predicted reliably [38–44] by ab-initio methods such as Density Functional Theory for Superconductors (SCDFT) [36, 37] or Eliashberg theory [47, 76], the latter also reviewed in section 2.2. However, as any such calculation relies on detailed knowledge of both electronic and phononic properties of a given material, the computational cost is large (in the order of weeks when performed on a single processor core).

It is therefore *unfeasible* to perform detailed calculations of T_c for each and every material during a high-throughput search for superconductors (chapter 4), where a *large* numbers of materials need to be tested for the desired property, i.e. T_c.

However, given that information about the superconducting properties of a material could be estimated from a set of quantities, available at low computational cost, this information can be employed both for discarding all materials where no superconductivity is expected to be found, and to provide a *ranking* in order to select the most promising materials, both in terms of probability and expected T_c, for more in-depth processing.

In this chapter, we propose a set of quantities within the reach of a ground state Kohn-Sham DFT calculation, termed *Descriptors of Superconductivity*:

1. Magnetic ordering (as an exclusion criterium) section 5.1
2. Density of states at Fermi level section 5.2
3. Fermi bond localization section 5.3
4. Group velocity at Fermi level (Fermi velocity) section 5.4

Filters and a ranking functions based on these descriptors are discussed in chapter 8 and chapter 10. All of these descriptors are based on theoretical considerations about the influence of electronic properties on superconductivity within Eliashberg theory (section 2.2), focusing mainly on two aspects:

a) properties which would be detrimental to superconductivity and therefore serve as trivial exclusion criteria

b) features of the electronic structure leading to large *isotropic* electron-phonon coupling strength λ,

as discussed in subsection 2.3.1, the critical temperature T_c is monotonically rising with λ, and it can be therefore safely considered the most influential parameter on superconductivity. Numerous such relations were evaluated in the process, however, in order to

simplify our prediction model, only the ones considered the most relevant entered the final model and are described within this work.

The selection of the *relevant* descriptors is the result of an iterative process, performed on a set of superconductors and non-superconductors (section 10.1), the latter definition including materials with insignificant T_c; starting from an initial set of materials with known T_c, a superconductivity prediction scheme was determined. Given the scheme, predictions of new superconductors and non-superconductors were made, evaluating the descriptors for a huge library of materials (chapter 7, chapter 8). These predictions were then verified by computationally expensive *ab-initio* calculations of the superconducting properties on a subset of predicted superconductors. In order to assess false predictions of superconductors as non-superconductors, the same was done for a few predicted non-superconductors. Prediction errors were then taken into account in order to propose new descriptors and a new scheme on the set of materials, now including also all new ab-initio superconductivity data. The idea of descriptors for superconductivity can be seen as a theoretical quantitative counterpart of the well known *Matthias rules*, e.g. summarized in [102, 103], which were a set of qualitative rules applied in experimental research. Matthias, who has experimentally discovered more materials with superconducting properties than any other researcher, found certain chemical and structural properties [104–107] which lead to the highest T_cs known at that time (note that many of those discoveries were made before BCS theory [46] provided a microscopic explanation for superconductivity[1]):

1. Transition metals are better than simple metals.
2. There are favorable electron/atom ratios (DOS_F peaks).
3. High symmetry is good; cubic symmetry is best.
4. Stay away from oxygen.
5. Stay away from magnetism.
6. Stay away from insulating phases.

Many materials with exceptional superconducting properties do not fulfill this set of rules: MgB_2, the conventional superconductor with the highest known T_c, has hexagonal symmetry, while the whole class of cuprate high-T_c superconductors *does* contain oxygen, is close to antiferromagnetic instability and derived from insulating parent compounds. In section 10.3, we will revisit some of these rules and relate them to our descriptors.

5.1. Magnetic order

As discussed in chapter 2, superconductivity and magnetism can only coexist under very restricted circumstances, neither of large interest in a high-throughput search for

[1] later, connections between rules and BCS were found [108–110]

superconductors with significant T_c; therefore, as a first filtering step, we determine the magnetic properties of each material in order to exclude the ones exhibiting magnetically ordered ground states.

There may be a few, albeit exotic, materials exhibiting *both* (triplet) superconductivity and ferromagnetism, such as $\mathsf{UGe_2}$ [111] or URhGe [112]. All systems belonging to this class show very low ($T_c < 1$K) superconducting transition temperatures. Superconductivity under such circumstances could only arise from triplet pairing, or from disjoint segments of the Fermi surface, some of them spin-degenerate and singlet-superconducting, others spin-polarized. Both low expected T_c and low frequency of the latter configuration justify an exclusion of any magnetically ordered system in our high-throughput search.

In order to setup our filter, we evalute the total spin polarization

$$m_{\mathrm{abs}}^{\mathrm{tot}} = \left| \sum_{j=1}^{\infty} f(\epsilon_j^{\uparrow}) \left\langle \psi_j^{\uparrow} \middle| \psi_j^{\uparrow} \right\rangle - \sum_{j=1}^{\infty} f(\epsilon_j^{\downarrow}) \left\langle \psi_j^{\downarrow} \middle| \psi_j^{\downarrow} \right\rangle \right|, \tag{5.1}$$

and the spin polarization projected on each atom a in the primitive cell

$$m_{\mathrm{abs}}^{a} = \left| \sum_{j=1}^{\infty} f(\epsilon_j^{\uparrow}) \sum_{i=1}^{N_{\mathrm{val}}} \left\langle \psi_j^{\uparrow} \middle| \phi_i^{a} \right\rangle - \sum_{j=1}^{\infty} f(\epsilon_j^{\downarrow}) \sum_{i=1}^{N_{\mathrm{val}}} \left\langle \psi_j^{\downarrow} \middle| \phi_i^{a} \right\rangle \right|, \tag{5.2}$$

where j runs over the KS states of the two spin channels in collinear LSDA, $f(\epsilon_j)$ denotes the corresponding occupation number and i enumerates the (pseudo-)atomic valence orbitals of atom a.

$$m_{\mathrm{abs}}^{\mathrm{max}} = \max_a m_{\mathrm{abs}}^{a}$$

is then taken as a measure for magnetism in a given material, indicating the maximum localized magnetic moment in a given material.

5.2. Density of States at Fermi Level

The relation between the density of states $\mathrm{DOS_F}$ and the isotropic electron-phonon coupling strength $\lambda \propto \mathrm{DOS_F}$ is a well-known concept already present in BCS theory, and has been reviewed in section 2.3.2. We therefore choose $\mathrm{DOS_F}$ as our first descriptor, which coincides with Matthias' second rule.

In numerical calculations, finite $\mathrm{DOS_F}$ may be observed in narrow-gap insulators (or semiconductors), where by definition $\mathrm{DOS_F} \equiv 0$, as the Fermi level lies within the band gap. This error is related to the fact that the smearing function (1.17) may lead to spurious occupation of conduction bands even at zero temperature. From the perspective of the $\mathrm{DOS_F}$ calculation using a smearing for interpolation purposes, the interpolation then assumes the existance of states within the gap.

Therefore, during our high-throughput search, we detect the possible presence of a band gap separately, and $\mathrm{DOS_F}$ is explicitly counted as zero in such cases. We employ

a simple band counting scheme for the task, and classify any material as an insulator, where the number of Kohn-Sham bands below E_F on all $\boldsymbol{k}$ points in our BZ sampling is identical.

5.3. Localization of bonds at Fermi energy

As mentioned in section 2.3.2, the contributions to λ can be decomposed into three parts: (1) the density of states at Fermi level, which covers the electronic phase space for electron-phonon scattering events, and which we treat explicitly as a descriptor (section 5.2) (2) the scattering amplitude for states at Fermi level subject to unit ion displacement (the deformation potential, section 2.3.2), independent of ionic mass and phonon frequency and (3) the amplitude of ionic motion, corresponding to the zero-point amplitude of a quantum harmonic oscillator and therefore incorporating the phonon frequency and the nuclear mass. In this section, we introduce the *Fermi bond localization* as a descriptor of superconductivity, providing an estimate for contribution (2); in the first part of this section, we demonstrate that this term gets larger, the more strongly localized the electronic wave function is in real space. In the second subsection, our method to *quantitatively* evaluate the degree of bond localization is presented.

The descriptor takes some inspiration from [113], where An and Pickett relate the strong electron-phonon coupling in $\mathsf{MgB_2}$ to metallicity of the covalent, localized B-B bonds, i.e. the presence of holes in those bonds. In a later publication [40], Bersier *et al.* presented a comparison of the electronic and superconducting properties of the isostructural and isoelectronic materials CaBeSi and $\mathsf{MgB_2}$. Defying all similarities in the band structure, and the fact that $\mathrm{DOS_F}$ of CaBeSi is a factor of 2 larger than in $\mathsf{MgB_2}$, the former has a neglegible T_c (≈ 0.4K), while the latter is the conventional superconductor with the highest known transition temperatures (≈ 39K). The difference in T_c originates in a large difference of the coupling parameter λ (0.8 vs. 0.4), which the authors relate to a strong difference in the structure of the charge distribution of the σ-bond states: In $\mathsf{MgB_2}$, this partial charge is more *localized* than in the case CaBeSi (Figure 5.6 illustrates this, as the Fermi charge is dominated by σ-bond character).

5.3.1. Bond localization and magnitude of the deformation potential

The reason for the large influence of the Fermi bond localization on the magnitude of the electron-phonon coupling is evident from the structure of the electron-phonon matrix elements

$$\tilde{g}^{n,n'}_{\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q},\nu} = \sqrt{\frac{\hbar}{2\Omega_{\boldsymbol{q},\nu}M}} \left\langle \boldsymbol{k}+\boldsymbol{q}, n' \middle| \Delta V^{\mathrm{scf}}_{\boldsymbol{q},\nu} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \middle| \boldsymbol{k}, n \right\rangle ,$$

which for Bloch states (1.14) read

$$= \sqrt{\frac{\hbar}{2\Omega_{\boldsymbol{q},\nu} M}} \int \mathrm{d}^3\boldsymbol{r}\, \mathrm{e}^{-i(\boldsymbol{k}+\boldsymbol{q})\cdot\boldsymbol{r}} u^*_{\boldsymbol{k}+\boldsymbol{q},n'}(\boldsymbol{r})\, \Delta V^{\mathrm{scf}}_{\boldsymbol{q},\nu}(\boldsymbol{r}) \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}\, \mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{r}} u_{\boldsymbol{k},n}(\boldsymbol{r})$$

$$= \sqrt{\frac{\hbar}{2\Omega_{\boldsymbol{q},\nu} M}} \int \mathrm{d}^3\boldsymbol{r}\, u^*_{\boldsymbol{k}+\boldsymbol{q},n'}(\boldsymbol{r})\, \Delta V^{\mathrm{scf}}_{\boldsymbol{q},\nu}(\boldsymbol{r})\, u_{\boldsymbol{k},n}(\boldsymbol{r}). \tag{5.3}$$

Given any perturbation $\Delta V^{\mathrm{scf}}_{\boldsymbol{q},\nu}(\boldsymbol{r})$ due to a phonon $\boldsymbol{q},\nu$, it is clear that the deformation potential contribution to the electron-phonon matrix element will rise with the *direct-space overlap* between potential perturbation $\Delta V^{\mathrm{scf}}_{\boldsymbol{q},\nu}(\boldsymbol{r})$, the source $u_{\boldsymbol{k},n}(\boldsymbol{r})$ and the target $u^*_{\boldsymbol{k}+\boldsymbol{q},n'}(\boldsymbol{r})$ elecronic states. For the electron-phonon matrix elements to be large, it is therefore a necessary condition to have a large overlap of the electronic wave functions with the region where the variation ΔV^{scf} is the strongest. In section 5.3.1, the variation will be shown to be maximal along the bond axis for bond stretching modes in a simple rigid-ion model, which in turn means that electronic states peaked along that axis are subject to stronger electron-phonon interaction. In the Kohn-Sham system, the self-consistent potential itself depends on the electronic density, so a second important contribution of the direct space charge distribution also enters the perturbation ΔV^{scf}.

An argument for focussing on the bonds can be derived from (5.3): Bloch states with a vanishing overlap of the lattice-periodic parts $u_{\boldsymbol{k}+\boldsymbol{q},n'}(\boldsymbol{r})$ and $u_{\boldsymbol{k},n}(\boldsymbol{r})$ must lead to vanishing matrix elements. First of all, this fact excludes transitions between states strongly localized at different sites, such as wave functions with a strong atomic character. Note that most of these states are already excluded by the restriction of our observations to states close to Fermi level; the eigenvalues of states counted as "core states", however, lie far lower in the energy range and are even considered as frozen in the pseudopotential approximation.

As conduction electrons, i.e. those close to the Fermi level, are orthogonal to the core states, they do not overlap with the core region. Therefore it is important to have strong ΔV^{scf} outside that region; this means that a bonding structure, where electrons (and so the Kohn-Sham potential) are localized, is favourable for large electron-phonon coupling.

This qualitative picture can be made more rigorous by constructing a simple model that we will discuss in detail in the following.

Model: rigid-ion chain

The effect of the localization of bond charge on the electron-phonon coupling can be demonstrated in a simple model: consider an chain of atoms aligned with the cartesian z-axis (Fig. 5.1, upper panel), leading to an external potential

$$V(\boldsymbol{r}) = \sum_J v_J(\boldsymbol{r})$$

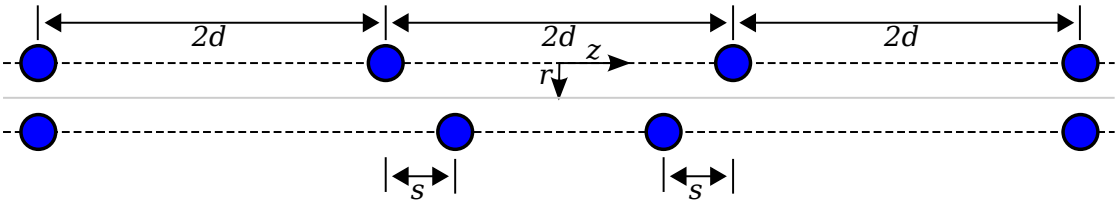


Figure 5.1.: Upper panel: Our model is an infinite chain of identical atoms, distance $2 \cdot d$. **Lower panel:** We evaluate the change in the potential when the 2 atoms closest to the origin are moved by s, simulating a long-wavelength bond-stretching mode.

composed of the ionic potentials $v_J(\boldsymbol{r})$, which we assume to be radial symmetric and identical, centered at the ionic sites:

$$v_J(\boldsymbol{r}) := v(|\boldsymbol{r} - \boldsymbol{R}_J|) \text{ with } \boldsymbol{R}_J = (2J-1)d\boldsymbol{e}_z$$

In the later steps, we will consider only longitudinal deformations: for purely transversal modes, the electron-phonon interaction would vanish [61]. Furthermore, in many superconductors, such as MgB_2, bond-stretching phonon modes play a dominant role in the total λ.

With this restriction in mind, we can express the external potential in cylinder coordinates (r, z, ϕ):

$$V(\boldsymbol{r}) = \sum_J v(\sqrt{r^2 + (z - (2J-1)d)^2 + \eta}),$$

where we included a small constant $0 < \eta \ll 1$ to ensure analyticity in the following steps; using a finite value for this parameter in the Coulomb potential $v(x) = \frac{1}{|x|}$ is a common technique for modelling a *soft Coulomb potential*, where the short-range divergence is screened, but the long-range behaviour remains unchanged.

Now consider a longitudinal, long-wavelength phonon mode, which displaces the two atoms closest to the origin symmetrically by s (Fig. 5.1, lower panel). Assuming a rigid-ion model, where the potential of an individual ion does not change due to the deformation, the perturbation in the potential reads:

$$\begin{aligned}\Delta_s V(r,z) = \quad & v\left(\sqrt{r^2 + (z - (d+s))^2 + \eta}\right) - v\left(r_-\right) \\ & + v\left(\sqrt{r^2 + (z + (d+s))^2 + \eta}\right) - v\left(r_+\right) \\ \text{with } r_+ := \sqrt{r^2 + (z+d)^2 + \eta} & \text{ and } r_- := \sqrt{r^2 + (z-d)^2 + \eta}\end{aligned}$$

If we expand into first order of the displacement s, we obtain

$$\Delta_s V(r,z) \approx s \cdot \frac{z+d}{r_+} v'(r_+) - s \cdot \frac{z-d}{r_-} v'(r_-).$$

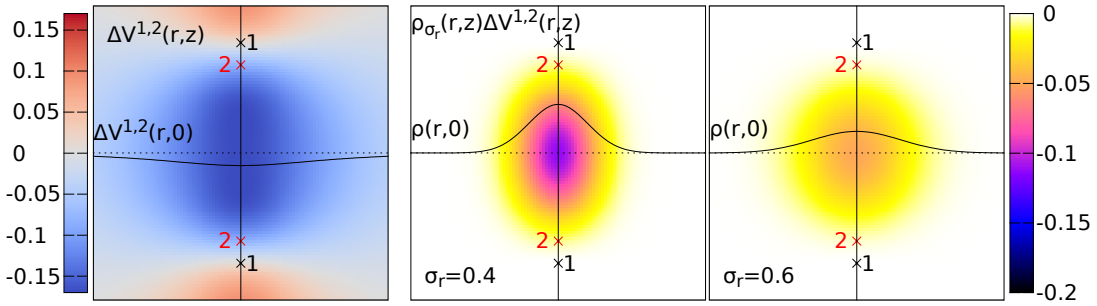


Figure 5.2.: Modelling the effect of bond localization when 2 atoms are displaced from position 1 to 2. **Left:** $\Delta V(r,z)^{\text{soft}}$ (5.4), black curve corresponds to $\Delta V(r,z)^{\text{soft}}$ for $z = 0$ **Middle/Right:** Integrand of (5.5) for states with different radial Gaussian widths, black curves correspond to $\rho^{\text{gauss}}_{\sigma_r}(r,z)$ for $z = 0$. Colorscales and $\sigma_z = 0.6$ are the same in both figures.

Note that the expansion is undefined at the original positions of the nuclei ($r = 0; z = \pm d$) when $\eta = 0$. The derivative with respect to r of this expression shows that the necessary condition for stationarity of the perturbation is fulfilled at $r = 0$, i.e. all points on the bond axis:

$$\frac{\partial \Delta_s V(r,z)}{\partial r} = s \cdot r \cdot \left[\frac{z-d}{r_-^2} \left(\frac{v'(r_-)}{r_-} - v''(r_-) \right) - \frac{z+d}{r_+^2} \left(\frac{v'(r_+)}{r_+} - v''(r_+) \right) \right],$$

independent of the actual shape of $v(r)$. For the Coulomb potential $v(r) = -\frac{1}{|r|}$, also the sufficient condition is fulfilled, as

$$\left. \frac{\partial^2 \Delta_s V(r,z)}{\partial r^2} \right|_{r=0} = -3 \cdot s \cdot \left(\frac{d-z}{\sqrt{\eta + (d-z)^2}^5} + \frac{d+z}{\sqrt{\eta + (d+z)^2}^5} \right) \neq 0,$$

as long as $d > 0$, which is fulfilled by construction. Therefore, the deformation potential for bond-stretching deformations

$$D^{\text{stretch}}(\boldsymbol{k}, n) = \left\langle \boldsymbol{k}n \middle| \Delta V^{\text{stretch}} \middle| \boldsymbol{k}n \right\rangle$$

of a state $|\boldsymbol{k}n\rangle$ is rising with the level of localization of $|\boldsymbol{k}n\rangle$ at the bond axis, due to a larger overlap with the region exhibiting the strongest perturbation of the potential.

We demonstrate now the effect in a simple model: consider an ionic *soft-Coulomb* potential (corresponding to a conventional Coulomb potential, when setting the earlier introduced analyticity parameter $\eta := 1$)

$$v(x) = -\frac{1}{\sqrt{x^2+1}},$$

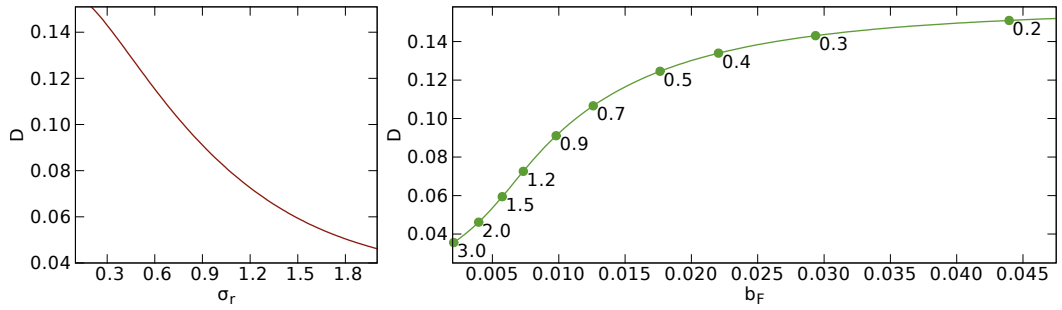


Figure 5.3.: Transistion matrix elements for unit bond-stretching deformations in our model. Perturbation and σ_z are kept fixed, while the radial Gaussian standard deviation is varied. **Left panel:** D_s displayed with σ_r as the independent variable. **Right panel:** D_s displayed with b_F from (5.10) as the independent variable. Numeric labels represent the corresponding values of σ_z.

which we employ to simulate the screening effect of the lower-lying states. The explicit form of the bare perturbation reads

$$\Delta_s V(r,z)^{\mathrm{soft}} = -\frac{s(d-z)}{\left((d-z)^2+r^2+1\right)^{3/2}} - \frac{s(d+z)}{\left((d+z)^2+r^2+1\right)^{3/2}}. \tag{5.4}$$

An example is presented in the left panel of Figure 5.2, where also the stationarity at $r=0$ for each z can be observed. As a second ingredient to our model, we assume that there are two Bloch states at Fermi energy, which differ only in the phase factor

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}u(r,z) \qquad \psi_{\boldsymbol{k}+\boldsymbol{q}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{q})\cdot\boldsymbol{r}}u(r,z)$$

and yield Gaussian densities (normalized to the unit cell)

$$\begin{aligned}\rho_{\sigma_r\sigma_z}(r,z) &= \langle\psi_{\boldsymbol{k}}|\widehat{n}(\boldsymbol{r})|\psi_{\boldsymbol{k}}\rangle = \left\langle\psi_{\boldsymbol{k}+\boldsymbol{q}}\middle|\widehat{n}(\boldsymbol{r})\middle|\psi_{\boldsymbol{k}+\boldsymbol{q}}\right\rangle = u^*(r,z)u(r,z)\\ &= \frac{1}{\sqrt{2\pi}^3\sigma_r^2\sigma_z}\exp\left(-\frac{r^2}{2\sigma_r^2}-\frac{z^2}{2\sigma_z^2}\right).\end{aligned}$$

For two states with a given bond localization, the bond-stretching deformation potential reads

$$\begin{aligned}D(\sigma_r) &= \left\langle\boldsymbol{k}+\boldsymbol{q}\middle|\Delta_s V(r,z)\mathrm{e}^{i\boldsymbol{q}\cdot\boldsymbol{r}}\middle|\boldsymbol{k}\right\rangle = \int u^*(\boldsymbol{r})\Delta_s V(r,z)u(\boldsymbol{r})\,\mathrm{d}^3\boldsymbol{r}\\ &= \int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{2\pi}\rho_{\sigma_r\sigma_z}(r,z)\Delta_s V(r,z)\;r\mathrm{d}\varphi\mathrm{d}r\mathrm{d}z.\end{aligned} \tag{5.5}$$

The middle and right panel of Fig. 5.2 display the integrand of (5.5), visualizing the effect of two different radial standard deviations. Model calculations have been performed, keeping σ_z, d and s fixed, while varying σ_r in order to simulate Fermi bonds with different real-space localizations, the result is presented in Figure 5.3. In the left

panel for small widths of the charge distribution an approximately linear decrease of $D(\sigma_r)$ can be observed; for larger widths, a saturation effect occurs, as the charge distribution approaches homogenity. The right panel of Figure 5.3 displays the relation of the model transition matrix elements and the corresponding *model-independent* measure b_F, the Fermi-bond localization measured within the Bader surface (5.10), which will be introduced in the following subsection 5.3.2. A strictly monotonic increase of the matrix element with increasing b_F can be observed, the slope first increasing in the low-b_F regime, while decaying above an inflection point around $\sigma_r = 1.5$. The shape of the curve has its origin in the shape of the states, and changes little (except for scale) when transitions occur induced by different perturbations.

5.3.2. A well-defined quantifier for the Fermi bond localization

The analysis presented in [40] is not directly applicable in the context of a high-throughput method, as it is based on a visual comparison of the charge localization of two structurally and electronically very similar materials. Especially, the character of the bands at Fermi level in both materials is comparable, a fact that allowed the authors to compare the partial densities originating from the same bond character.

In this section, we describe our idea to generalize the aforementioned idea to a quantitative "Fermi bond localization".

Our approach consists of an analysis of the Fermi charge distribution, quantifying the corresponding bond localization by integrating the absolute value of the charge gradient $|\nabla\rho(\boldsymbol{r})|$ in the direct-space region relevant for bonding: this quantity will be large when strong localization occurs, and assuming a homogenous charge distribution, corresponding to metallic bonding, it will vanish. One difficulty in the introduction of a bond-related quantity lies in the separation of "ion" and "bond" states, due to a missing strict and *consistent* definition of such a separation, especially in the context of a plain-wave basis Kohn-Sham DFT method. Moreover, the charge varies rapidly in the core region and would easily dominate the result, so care must be taken to restrict the analysis to regions sufficiently far from the core. An approach by removing spherical regions around the nuclei proved problematic, due to the arbitrariness of the radii and the assumption of a spherical shape.

In order to achive the goal of consistent description, we turn to a technique known as *Bader analysis*, and restrict our analysis to the charge distribution within surfaces, rigorously defined by the total charge density.

Bader Surface

Bader formulated [114] a *Theory of Atoms in Molecules*; in this context, he suggested a particularly elegant scheme for partitioning space into volumes associated to the constituent atoms of a molecule, which can easily be generalized to periodic boundary conditions (Figure 5.4 illustrates this concept): the *total* charge density $\rho(\boldsymbol{r})$ of a material does always exhibit maxima close to the nuclei (due to the low-lying core states) and as a trivial consequence there exist density minima in the interstitial region. A *Bader surface*

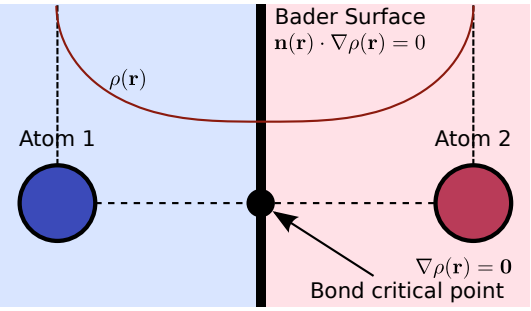


Figure 5.4.: Bader Surface and Bader Atomic Volumes

(Fig. 5.4, black vertical line) is a surface $\mathcal{B} \subset \mathbb{R}^3$ with normal vectors $\boldsymbol{n}(\boldsymbol{b}) : \mathcal{B} \to \mathbb{R}^3$, fulfilling

$$\boldsymbol{n}(\boldsymbol{b}) \cdot (\boldsymbol{\nabla}\rho(\boldsymbol{r})) \equiv 0 \quad \forall \boldsymbol{b} \in \mathcal{B}. \tag{5.6}$$

In words: the surface defined by consisting of points at charge-density minima with respect to lines perpendicular to the surface. The point $\boldsymbol{r}_\mathrm{c}$, where $\boldsymbol{\nabla}\rho(\boldsymbol{r}_\mathrm{c}) = \mathbf{0}$ is called the *bond critical point* in Bader's theory; the density $\rho(\boldsymbol{r}_\mathrm{c})$ at this particular point can be interpreted as a measure for the shared electron density within the bond, and thus be used to classify the bond as "ionic" (when $\rho(\boldsymbol{r}_\mathrm{c})$ is small) or "covalent" (when $\rho(\boldsymbol{r}_\mathrm{c})$ is large).

Conventionally, $\mathcal{B}$ is directly used for partitioning of direct space: the volume $\mathcal{V}_I$ (Fig. 5.4, colored backgrounds) between the nuclear equilibrium position of atom I and the segments of $\mathcal{B}$ enclosing it is defined as being associated with atom I. Evaluation of a quantity on $\mathcal{V}_I$ can than be performed in order to find atom I's contribution. A typical example is the evaluation of local charge

$$\rho_I = \int_{\mathcal{V}_I} \rho(\boldsymbol{r})\, \mathrm{d}^3\boldsymbol{r}$$

in order to evaluate the ionization state or spin polarization

$$m_I = \int_{\mathcal{V}_I} \rho^{\uparrow}(\boldsymbol{r}) - \rho^{\downarrow}(\boldsymbol{r})\, \mathrm{d}^3\boldsymbol{r}$$

for assigning a localized magnetic moment to a particular atom.

Bader Bond Localization

We follow a slightly different route, as we intend to quantify the localization of the Fermi bond charge. By its definition (5.6), the Bader surface lies in an area deemed most representative for the chemical bonds [114] of a given system, as it is constructed around the bond critical points (Fig. 5.5, left panel). The gradient of the total charge density $\boldsymbol{\nabla}\rho(\boldsymbol{b})$ within the Bader surface refers only to spacial variations within the surface, as $\boldsymbol{n}(\boldsymbol{b}) \cdot \boldsymbol{\nabla}\rho(\boldsymbol{b}) \equiv 0$. The localization of the bond charge between atoms I and J in

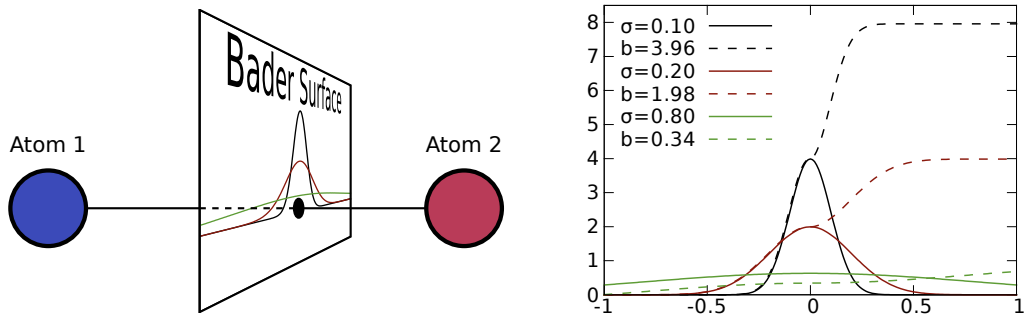


Figure 5.5.: Concept of the Bader bond localization b: $\rho(\boldsymbol{r})$ is evaluated within the Bader surface (left). The right figure demonstrates the evaluation of (5.8) for bond charges with different localization properties, described by Gaussian distributions on a closed 1d Bader surface (ring).

a molecule, with a (possibly infinitely large) separating surface segment $\mathcal{B}_{I,J}$, can be evaluated from

$$\tilde{b}^{I,J} = \int_{\mathcal{B}_{I,J}} \rho(\boldsymbol{b}) |\boldsymbol{\nabla}\rho(\boldsymbol{b})| \, \mathrm{d}B / \int_{\mathcal{B}_{I,J}} \rho(\boldsymbol{b}) \, \mathrm{d}B.$$

The localization of all bonds in a system can then be determined as

$$\tilde{b} = \int_{\mathcal{B}} \rho(\boldsymbol{b}) |\boldsymbol{\nabla}\rho(\boldsymbol{b})| \, \mathrm{d}B / \int_{\mathcal{B}} \rho(\boldsymbol{b}) \, \mathrm{d}B. \tag{5.7}$$

Restricting to the case of solids (where all atomic volumes $\mathcal{V}_I$ must be closed) and restricting the integration to the Bader surface within the unit cell $\mathcal{B}_\mathrm{C}$,

$$b = \int_{\mathcal{B}_\mathrm{C}} |\boldsymbol{\nabla}\rho(\boldsymbol{b})| \, \mathrm{d}B / \int_{\mathcal{B}_\mathrm{C}} \mathrm{d}B \tag{5.8}$$

could be used as a measure for the total bond localization unbiased by the total charge within the surface, as the normalization can be performed with respect to the surface area in the case of closed volumes. The right panel of figure 5.5 demonstrates this concept on a two-dimensional model: in this case, a closed Bader surface corresponds to a ring around each atom. The Bader bond localization (5.8) was evalutated for different normalized (model) Gaussian charge distributions ranging from very localized ($\sigma = 0.1$) to almost homogenous ($\sigma = 0.8$), clearly demonstrating the effect of the localization properties on our b.

Fermi Bader Bond Localization

We have to introduce another important generalization over the analysis presented in Ref. [40] to obtain a quantity applicable to a wide range of materials: the authors presented a comparison of bond charges resolved by bond character, which was only possible due to the similarity of the two presented materials.

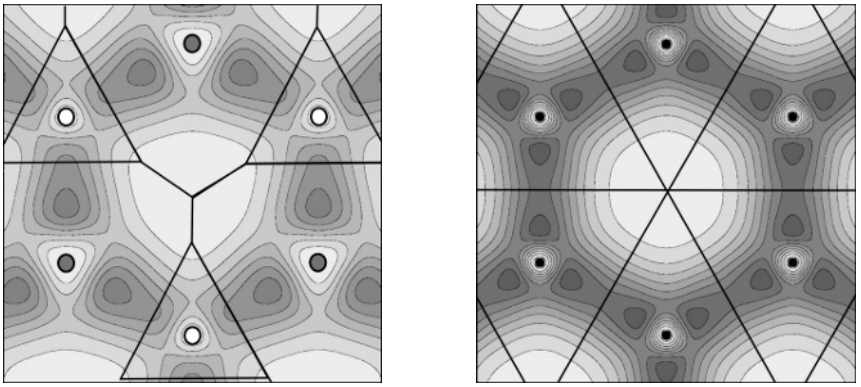


Figure 5.6.: Local density of states at Fermi level of CaBeSi (left) and MgB2 (right) in the hexagonal plane. Thick lines are schematic representations of the Bader surface (determined from the *total* charge). Color scales and contour levels are identical in both figures.

As a first step, based on the insight that only transitions between electronic states very close to Fermi level contribute to $\alpha^2F(\omega)$, we introduce the concept of a *Fermi local density of states*

$$N_{\mathrm{F}}(\boldsymbol{r}) = \sum_{n,\boldsymbol{k}} \langle \psi_{n\,\boldsymbol{k}} | \widehat{n}(\boldsymbol{r}) | \psi_{n\,\boldsymbol{k}} \rangle \, \delta(\epsilon_{n,\boldsymbol{k}} - E_{\mathrm{F}}), \tag{5.9}$$

corresponding to the charge density of all electronic states at Fermi level, if they were all fully occupied.

We can now evaluate (5.8) for $N_{\mathrm{F}}(\boldsymbol{r})$ instead of $\rho(\boldsymbol{r})$, in order to determine the localization of the bond states actually contributing to the electron-phonon coupling:

$$b_{\mathrm{F}} = \int_{\mathcal{B}_{\mathrm{C}}} |\boldsymbol{\nabla} N_{\mathrm{F}}(\boldsymbol{b})| \, \mathrm{d}B \Big/ \int_{\mathcal{B}} \mathrm{d}B, \tag{5.10}$$

which we include in our set of descriptors.

Figure 5.6 visualizes this concept for the 2 cases compared in [40], applied within the BeSi layers of CaBeSi and the B_2 layers of MgB_2. Both density plots display $N_{\mathrm{F}}(\boldsymbol{r})$ within the layers, while the gradients can be read from the density of the contour lines (color maps and contour levels are identical in both subgraphs in order to ease a visual comparison). Schematic representations of the Bader surface[2] intersections with the layers are drawn as thick black lines. Based on the qualitative comparison, a significantly larger b_{F} (5.10) is expected in MgB_2 than in CaBeSi.

Numerical approximation of the surface integral In practical calculations, quantities such as $\rho(\boldsymbol{r})$ or $\rho_F(\boldsymbol{r})$ are represented on discrete grids $\mathcal{R} = \{\boldsymbol{r_i}\}$ within the unit cell of the direct lattice.

[2] determined via the method in A.1

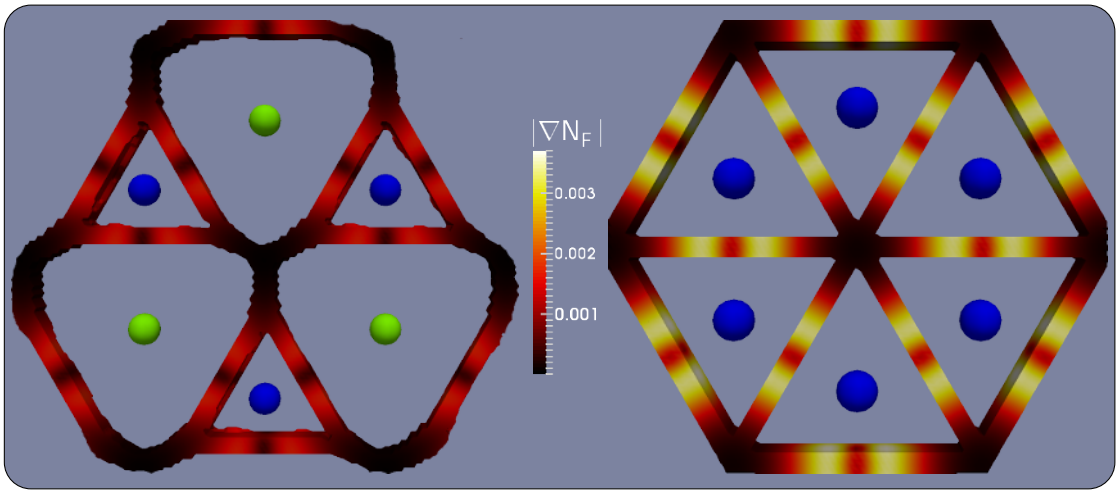


Figure 5.7.: Fermi Bader Bond Localization in our implementation: $|\boldsymbol{\nabla} N_{\mathrm{F}}(\tilde{\boldsymbol{b}})|$ in CaBeSi (left) and MgB_2 (right). $\tilde{\boldsymbol{b}}$ are grid points in the vincinity of the Bader surface, Bader surfaces in both figures are displayed only in the vincinity of the BeSi rsp. B_2 planes. Color scales are identical in both figures.

Therefore we approximate the surface integral in (5.10) by a volume integral

$$b_{\mathrm{F}} \approx \lim_{\varepsilon \to 0} \int_{\mathcal{B}_\varepsilon} |\boldsymbol{\nabla} N_{\mathrm{F}}(\boldsymbol{r})| \mathrm{d}V / \int_{\mathcal{B}_\varepsilon} \mathrm{d}V$$

in a small shell around the Bader surface $\mathcal{B}$:

$$\min_{\boldsymbol{b} \in \mathcal{B}} |\boldsymbol{n}(\boldsymbol{b}) \cdot (\boldsymbol{b}_\varepsilon - \boldsymbol{b})| \leq \varepsilon \quad \forall\, \boldsymbol{b}_\varepsilon \in \mathcal{B}_\varepsilon.$$

Assuming that, based on

$$\begin{aligned} \boldsymbol{n}(\boldsymbol{r}_{\mathrm{B}}) \cdot \boldsymbol{\nabla} \rho(\boldsymbol{r}_{\mathrm{B}}) &= 0, \\ \text{also } \boldsymbol{n}(\boldsymbol{r}_{\mathrm{B}}) \cdot \boldsymbol{\nabla} N_{\mathrm{F}}(\boldsymbol{r}_{\mathrm{B}}) &\ll |\boldsymbol{\nabla} N_{\mathrm{F}}(\boldsymbol{r}_{\mathrm{B}})|, \end{aligned}$$

b_{F} is only weakly dependent on ε. With this assumption in mind, the value of ε is chosen as a small multiple of the distance between grid points in direct space, while compensating for grid anisotropy.

Example: CaBeSi and MgB_2

We will now briefly review the results for our measure of the Fermi bond localization, on the two example materials CaBeSi and MgB_2 mentioned earlier in this section, in order to illustrate the application of the method. As described in section A.1, the surface integral is replaced by volume integration within a thin shell of width ε around the Bader surface.

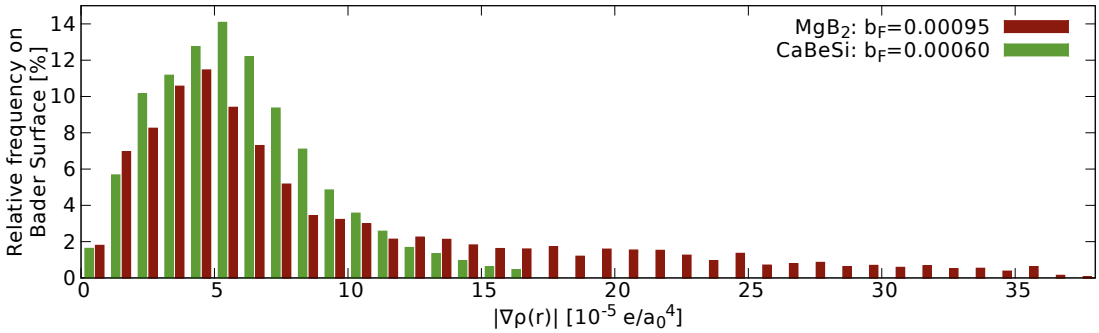


Figure 5.8.: Normalized statistical distribution of $|\boldsymbol{\nabla}\rho(\boldsymbol{r}_{\mathcal{B}})|$ in CaBeSi and MgB_2

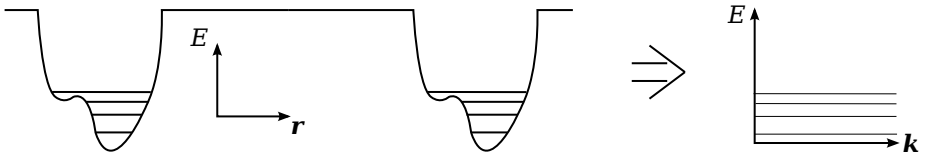


Figure 5.9.: A strictly molecular solid corresponds to a narrow-band insulator

Figure 5.7 displays the integrand $|\boldsymbol{\nabla}N_{\mathrm{F}}(\boldsymbol{r})|$ (with identical color maps) on the integration volume $\mathcal{B}_\varepsilon$. Only a slice of $\mathcal{B}_\varepsilon$ in the vincinity of the BeSi (B_2) planes of CaBeSi (MgB_2) is shown in the figures. While the Fermi charge around the surface in CaBeSi exceeds the one in MgB_2 by about 50%, the large difference in charge localization can be clearly observed from the magnitude of the charge gradient used to color both surfaces.

The electronic states around the Fermi level contribute to $N_{\mathrm{F}}(\boldsymbol{r})$ mostly within this region surrounding the planes, while little Fermi charge can be found in the surface segments between the plane atoms and the alkaline earth atoms above/below the planes. This fact is reflected in a large baseline of comparatively low $|\nabla N_{\mathrm{F}}|$ values in both cases, displayed as normalized histograms in Figure 5.8.

Therefore, our measure for the bond localization yields an increase by 60% in MgB_2, compared to CaBeSi.

5.4. Fermi velocity

There is, however, a whole class of systems consisting of crystals with what could be called a strong molecular character, where even strongly localized Fermi bonds would not contribute significantly to the global electron-phonon coupling λ, if the system was indeed metallic.

Based on this idea, we define a *molecular solid* as a system S, which consists of finite local subsystems $s_1, s_2, \ldots, s_n$, which behave like isolated molecules; bonds with covalent character are only formed locally among the members of each s_i, and the extent of electronic valence density into the intermediate regions is neglegible (Fig. 5.9, left panel).

Effects of the lattice periodicity on the electronic states, which normally lead to a

significant broadening of the electronic eigenvalues into energy bands with a finite dispersion, are small in all valence bands of such a system, i.e. the group velocity

$$\boldsymbol{v}_{\boldsymbol{k}\,n} = \frac{1}{\hbar}\frac{\partial \epsilon_{\boldsymbol{k}\,n}}{\partial \boldsymbol{k}} \tag{5.11}$$

of all valence states is vanishingly small. Therfore, the spectrum is composed of sharp δ peaks of the isolated subsystems' *molecular* electronic states, in principle making such a system an insulator with a band gap too large for phonon-induced transitions (Fig. 5.9, right panel). A single phonon carries an energy of around $10 - 100\text{meV}$, which can be absorbed or emitted during a transition between electronic states. In the case of an insulator, the band gap lies in the order of 1eV, which is outside the range of phonon-induced transitions. Therefore only metals can exhibit superconductivity, and the phonon energy scale is also the reason why our research is focused on the Fermi energy.

In our high-throughput calculations, we are restricted to computationally cheap Kohn-Sham DFT exchange-correlation functionals such as local spin density approximation (LSDA), which may predict a metallic ground state in such systems. A fundamental reason, in the context of molecular solids, is the poor description of strong correlation effects by LSDA, such as in the prototypical failure to describe the transition to a paramagnetic, insulating state when increasing the interatomic distance in a lattice of hydrogen atoms [115]. Members of this class of systems, where insulating behaviour arises from correlation effects, are called *Mott-Hubbard insulators*.

Moreover, in the case of very low electronic bandwidth and therefore low Fermi velocity v_F, Eliashberg theory of superconductivity (section 2.2) becomes invalid: it is founded on the Migdal theorem, which on the basis of an adiabatic parameter $\propto v_\mathrm{F}^{-1}$ justifies the neglection of all phononic contributions to the electron self energy containing more than one phonon propagator (vertex corrections). Obviously, this assumtion is violated in systems where v_F is small.

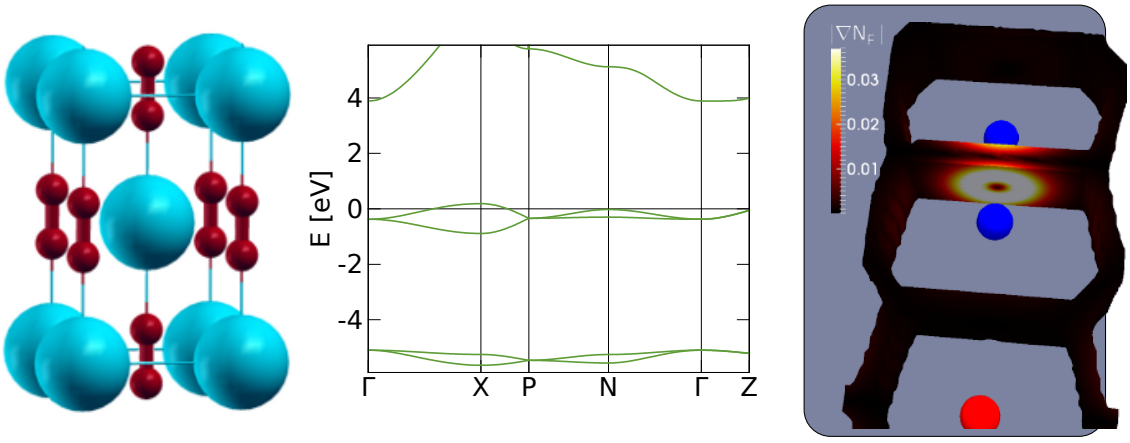
Despite the LSDA's fundamental unreliability in the context of strongly correlated systems, the KS bands in the case under consideration, i.e. molecular solids, tend to be narrow. Due to this fact, we can use the group velocity of states at the Fermi level to estimate in howfar a given system exhibits features resembling a molecular solid, and introduce an isotropic measure

$$\overline{v_\mathrm{F}} := \sum_n \int |\boldsymbol{v}_{\boldsymbol{k}\,n}|\,\delta(\epsilon_{\boldsymbol{k}\,n} - E_\mathrm{F})\mathrm{d}\boldsymbol{k} \Big/ \sum_n \int \delta(\epsilon_{\boldsymbol{k}\,n} - E_\mathrm{F})\mathrm{d}\boldsymbol{k}, \tag{5.12}$$

as an additional descriptor of superconductivity, introduced in order to filter out cases where predictions with present theory can be problematic.

5.4.1. Example: KO_2

One example of such systems appearing in our high-throughput search is KO_2. As displayed in Figure 5.10a, the crystal structure can be described by O_2 dimers, embedded in a matrix of potassium atoms, with the potassium atom formally donating its valence electron to the dimer.



(a) Crystal structure: O_2 dimers, embedded in a potassium matrix

(b) LDA band structure, exhibiting very low v_F in the almost molecular O_2 π bands

(c) $|\nabla N_F(\boldsymbol{r})|$ within the Bader surface; Colormap is rescaled by factor 10, compared to Figure 5.7

Figure 5.10.: KO_2: a strongly correlated, molecular crystal, predicted falsely as metallic within LDA

In experiment [116] the O_2 dimers show properies resembling isolated molecules, and due to the $\frac{3}{4}$ filling of their π_g molecular orbitals, possess localized magnetic momenta. This fact gives rise to a complex magnetic phase diagram of the material, with an antiferromagnetic phase at temperatures below 7.1 K. Moreover, KO_2 is an insulator.

Our simulations predict the system to be metallic and nonmagnetic, due to a couple of reasons. For one, as localized magnetic momenta are a feature hardly observed within *sp* systems, we do not include spin degrees of freedom when performing calculations on them, to lower the computational cost. As this system is *strongly correlated*, also simple LDA and GGA calculations including spin do not yield results in agreement with experiment, which could, in principle, be achieved by more complex approaches, such as ,,generalized Kohn-Sham" [117, 118].

The electronic charge is strongly localized within the π bonds of the oxygen dimers; actually, the resulting Fermi bond localization, computed via the method described previously, is within the highest 1% observed on any material in our high-throughput search.

However, due to the isolated character of the O_2 dimers, such states exhibit little dispersion (in the band structure picture), a fact that is reflected in the Fermi velocity v_F, belonging to the lowest 6% of all materials predicted as metallic by our LSDA calculations. Therefore, in the example of KO_2, v_F provides sufficient information to classify the system as molecular and strongly correlated.

5.5. Summary

In this chapter, the structure of a computational experiment to be performed on each of the candidate materials in a high-throughput search for superconductors has been proposed, which consists of the numerical evaluation of set of scalar quantities, the *descriptors of superconductivity*:

- absolute magnetization $m_{\mathrm{abs}}^{\mathrm{tot}}$ and spin polarization per atom m_{abs}^{a}
- density of states at Fermi level $\mathrm{DOS}_{\mathrm{F}}$
- Fermi bond localization, evaluated within the Bader surface b_{F}
- isotropic Fermi velocity $\bar{v}_{\mathrm{F}}$

All these quantities can be evaluated within the framework of Kohn-Sham density functional theory, and are well founded on theoretical considerations, supported by empirical data. The computational cost of each experiment is moderate, ranging from a processorminutes to a few processorhours depending on the size of the system; a small-scale high-throughput search therefore can be very well performed with the help of a present computation facility.

In chapter 8, we present statistical data on each descriptor gathered during our high-throughput search for superconductors, performed on approximately 8.000 different materials. As discussed within the present section, our descriptors are expected to show some degree of interdependence; chapter 8 therefore includes a quantitative analysis on the latter. A simple prediction model for superconductivity based on $\mathrm{DOS}_{\mathrm{F}}$, b_{F} and v_{F} will be introduced and evaluated in chapter 10, establishing a relative weighting of the quantities against each other.

In the next chapter, we will describe our methodological development to apply machine learning techniques to the prediction of such descriptors directly from the crystal structure.

6. Representation of crystal structures for machine learning

In the previous chapter, we introduced our descriptors of superconductivity in order to obtain a rough estimate of the superconducting properties within Kohn-Sham DFT at a moderate computational cost.

However, this moderate cost still poses a serious bottleneck for high-throughput methods (HTMs): raising the limits on the unit cell sizes and chemical compositions, the thereby defined subspace of chemical compound space $\mathcal{C}$ becomes so large, and the complexity of the unit cells so high, that, even within the efficient framework of Kohn-Sham density functional theory (KS-DFT, subsection 1.1.2), a systematic high-throughput exploration grows beyond reach for present-day computing facilities.

Recently, machine learning methods (ML, chapter 3) have contributed accurate models for predicting molecular properties [92, 93], transition states [119], reaction surfaces [120], potentials [121] and self-consistent solutions for DFT [122]. All these applications deal with finite systems (atoms, molecules, clusters).

Extending such methods to *periodic* systems would provide a strong boost to the throughput in a high-throughput search for superconductors, given that such methods could be applied to directly predict the descriptors from the crystal structure of a given material $\boldsymbol{m}_i \in \mathcal{C}$: while a typical KS-DFT calculation of our descriptors of superconductivity (chapter 5) may require a couple of minutes up to a few hours to complete, ML methods typically only require fractions of a second for a single prediction.

Discrete-valued descriptors, such as the metallicity of a given material, map to *classification problems* in machine learning terminology, and can be predicted by a support vector machine (SVM, subsection 3.4.1). Continuous-valued descriptors (e.g. the Fermi density of states) pose a *regression problem*, to which *kernel ridge regression* (KRR, section 3.4.2) is applied.

Both techniques assume nonlinear maps between input data (representations of the crystal structure) and the observable being predicted. Whether or not this unknown map can be approximated by the predictor depends strongly on the representation chosen for the input [123–125].

To be more specific: in order to employ machine learning methods, one needs to represent all crystal structures as vectors $\boldsymbol{x}_i$ in an input (Hilbert) space $\mathcal{X}$, i.e. establish a map $x : \mathcal{C} \to \mathcal{X}$. The nonlinear problem is linearized by mapping from $\mathcal{X}$ into a potentially high-dimensional feature (Hilbert) space $\mathcal{F}$ (section 3.3). Given that a linear algorithm's dependence on the input vectors $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ can be expressed solely by inner products $\langle \boldsymbol{x}_i | \boldsymbol{x}_j \rangle$, mapping to $\mathcal{F}$ can be implicitly performed by the substitution of the inner product by an appropriate kernel function (subsection 3.3.1). In the context of

this work, radial basis function (RBF) kernels, such as the Gauss kernel, are employed. Such kernels do only depend on the distance between the representations $\left|\boldsymbol{x}_i - \boldsymbol{x}_j\right|$ in input space.

Therefore, the central task of extending ML techniques towards our field of endeavour consists in finding a representation and distance measure that quantitatively describes the similarity, considering the label to be predicted, of any two given crystal structures $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

For finite systems, one particular way of representing finite systems, namely the so-called Coulomb matrix (CM), has been very successful [93]; we outlined this representation in section 3.5.

In section 6.2, we briefly discuss fundamental problems of extending the Coulomb matrix representation to crystal structures by information contained in the Bravais vectors (B+CM).

The thereby gathered insight is incorporated in a novel statistical representation of crystal structures, the *partial radial distribution function* (PRDF), presented in section 6.3, with significantly enhanced prediction performance over any B+CM approach.

This chapter documents our collaboration with the Machine Learning Group at TU Berlin [126].

6.1. Conventional description of crystal structures

In the solid state community, crystals are conventionally described by the combination of the *Bravais Matrix*, containing the primitive translation vectors $\boldsymbol{a}_{1\ldots3}$, and the *basis*, setting the position $\boldsymbol{R}_I$ and type Z_I of the atoms in the unit cell. This type of description is not unique and thus not a suitable representation for the learning process since it depends on a number of arbitrary choices: first of all, the coordinate system for both $\boldsymbol{a}_i$ and $\boldsymbol{R}_I$ may be rotated or translated, which, while leaving the actual crystal structure unchanged, leads to significant change of the cartesian components, and an artificial ordering is imposed on the basis atoms. While both former cases apply also to molecules, there is an additional degeneracy for crystal structures: the choice of the unit cell, as any structure $\boldsymbol{m}$ may also be described by any supercell $\tilde{\boldsymbol{m}}$, adapting both bravais vectors and basis. Summarizing, there exists an infinite number of equivalent representations that would be perceived as distinct crystals by the machine. In principle, recognizing equivalent representations could also be tackled by machine learning directly as done for molecules in Ref. [92, 127, 128]. However, a significant computational cost in terms of size of the training set had to be paid. Due to the aforementioned larger ambiguity in the case of crystals, an even higher cost is expected.

6.2. Problems in Coulomb-matrix-inspired representations of crystals

As the Coulomb matrix representation (section 3.5) proved valuable in the representation during ML experiments on molecules, it is tempting to to simply extend the representa-

tion by information about the lattice, a family of representations we summarize under the label "Coulomb matrix extended by Bravais vectors" (B+CM).

However, there are additional levels of degeneracy in conventional descriptions by atomic basis $\mathcal{B}$ and lattice vectors $\boldsymbol{a}_{1\ldots3}$ of crystal structures, besides the degeneracy arising from the choice of rotation and translation (as a whole) also present in molecules:

choice of lattice vectors and unit cell, as the same material can be also described by any supercell, e.g. basis $\tilde{\mathcal{B}} = \mathcal{B} \cup \{\mathcal{B} + \boldsymbol{a}_1\}$ and lattice vectors $\tilde{\boldsymbol{a}}_1 = 2\boldsymbol{a}_1, \boldsymbol{a}_{2\ldots3}$.

translation of any basis atom by a direct lattice vector leaves the structure described is unchanged.

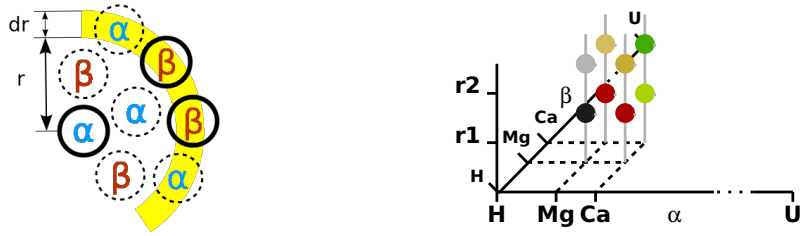
While the latter could be, in principle, compensated by employing $\min_{\boldsymbol{R}} |\boldsymbol{R}_i - \boldsymbol{R}_j - \boldsymbol{R}|$ as a distance measure in the off-diagonal denominator of a CM representation, the set of $\boldsymbol{R}$ available to the minimization depends on the former choice, causing further ambiguity in the description. Moreover, when the distance measure between any two structures with different cell sizes are computed, supercell expansions to the least common multiple of both cells have to be performed, as the sparse row/colums scheme employd in the molecular case would not account for the environment in an infinite, filled solid. The resulting comparatively large cell sizes yield a combinatorial nightmare, due to the factorial dependence of the possible atomic permutations (3.16) on the cell size. The same consideration is even more emphasized in the representation of crystals by CMs of finite, representative clusters. In principle, complexity could be reduced by a coarser sampling of the index permutations; however, this reduction comes with a significant cost in prediction accuracy.

Furthermore, as the bare nuclear charges are entering in the numerators, the distance measure can be interpreted as employing differences of nuclear charges as a measure of similarity between chemical elements, which strongly contradicts chemical insight. This fact may be less problematic in the case of [93], where a subspace of $\mathcal{C}$ is, by choice of the very restricted set of constituent elements and the size training set, densely sampled, effectively reducing the probability of information inferred from other elements. However, as the set of constituent elements in our dataset (chapter 8) spans most of the periodic table, while the number of crystal structures in ICSD is comparatively small, the sampling in our subspace of $\mathcal{C}$ is far coarser, raising the probability of such undesirable information transfer.

In summary, all B+CM representations in our experiments either turned out too complex for practical application or did not account for fundamental symmetries of periodic systems in a way suitable for SVM and KRR predictors. Therefore, while delivering a complete description of crystal structures (from a human point of view), the resulting accuracy of predictions is far worse than in the molecular case [93].

6.3. Partial Radial Distribution function

Based on these observations, we can now formulate a set of disirable properties of a representation of crystal structures employable for machine learning predictions of global



(a) Evaluating the density of β atoms in a single shell around α

(b) Feature matrix X representing a hypothetical MgCa alloy, sampled at two radial distances $r1$ and $r2$; α and β axis correspond to element pairs

Figure 6.1.: Partial radial distribution function, representing the density of atoms of type β in spherical shells around an atom of type α

properties, such as the density of states at Fermi level $\mathrm{DOS_F}$ and the metallicity. These properties can be summarized as "invariance to human choice" of the representation of a single crystal; the representation should be invariant to

- choice of the cartesian axis (global rotation and translation)
- choice of atom indices (any reordering of atoms)
- choice of the unit cell (crystallographic description by a supercell $\tilde{\boldsymbol{m}}_j$ of material $\boldsymbol{m}_j$ should map to an identical ML representation).

Furthermore, the distance measure should not derive similarity of inequal periodic elements from the difference of nuclear charge.

A very simple representation fulfilling could be vectors

$$\boldsymbol{n} = (n_{\mathsf{H}}, n_{\mathsf{He}}, \ldots, n_{\mathsf{U}})/V,$$

where n_{el} is the number of atoms of the corresponding element per unit cell, and V is the volume thereof. The information about periodic elements is such encoded by the index of component, and each component contains a density of nuclei of the given type. However, in this representation, effectively all information about the geometry of the crystal structure is lost.

This loss of geometric information is especially severe, as the physical properties of a material depend strongly on the formation of chemical bonds between a given atom and the neighbouring ones, where, besides the elements involed, also the interatomic distance plays an important role.

In order to retain information, both on bond lengths and on the overall crystal structure, we employ an approach to represent a crystal structure by *pair distribution functions*: given any two atoms I and J of the basis, the density of (potentially periodic replica of) J in a spherical shell at a distance r and of width dr around atom I is given

by

$$\tilde{g}^{dr}_{IJ}(r) := \frac{1}{V^{dr}(r)} \sum_{\boldsymbol{R}} \theta(|\boldsymbol{R}_I - \boldsymbol{R}_J + \boldsymbol{R}| - r)\theta(r + dr - |\boldsymbol{R}_I - \boldsymbol{R}_J + \boldsymbol{R}|)$$

where $\boldsymbol{R}$ are direct lattice vectors, and

$V^{dr}(r) := \frac{4}{3}\pi((r + dr)^3 - (r^3))$

is the volume of the associated spherical shell. In this form, a pair distribution function is not applicable as machine learning input, as indices I and J run over the basis, which changes from material to material. However, we may switch the indices to enumerate chemical elements, and normalize with respect to the number of sites in the unit cell occupied by the first element in order to regain unit cell choice invariance:

$$\begin{aligned} g^{dr}_{\alpha\beta}(r) := & \frac{1}{N_\alpha} \sum_I \frac{1}{V^{dr}(r)} \sum_J \sum_{\boldsymbol{R}} \\ & \theta(|\boldsymbol{R}_I - \boldsymbol{R}_J + \boldsymbol{R}| - r)\theta(r + dr - |\boldsymbol{R}_I - \boldsymbol{R}_J + \boldsymbol{R}|)\delta_{Z_\alpha, Z_I}\delta_{Z_\beta, Z_J} \\ = & \frac{1}{N_\alpha V^{dr}(r)} \sum_{i=1}^{N_\alpha} \sum_{j=1}^{N_\beta} \sum_{\boldsymbol{R}} \theta(|\boldsymbol{R}_{\alpha_i} - \boldsymbol{R}_{\beta_j} + \boldsymbol{R}| - r)\theta(r + dr - |\boldsymbol{R}_{\alpha_i} - \boldsymbol{R}_{\beta_j} + \boldsymbol{R}|). \end{aligned}$$

We call this representation the *partial radial distribution function* (PRDF), which describes the density of atoms of type β around an atom of type α, averaged over all instances of α within the unit cell (Figure 6.1a). The distribution is globally valid due to the periodicity of the crystal and the normalization with respect to the considered crystal volume. The PRDF can be related to radial distribution functions as used in the physics of x-ray powder diffraction [129] and text mining from computer science [130, 131].

As input for the learning algorithm, given a material $\boldsymbol{m}$, we employ the feature matrix

$$X(\boldsymbol{m}) \text{ with entries } x_{\alpha\beta,n} = g^{dr}_{\alpha\beta}(\boldsymbol{m}, r_n) \tag{6.1}$$

i.e., the PRDF representation of all possible pairs of elements as well as shells at fixed set of radii up to an empirically chosen cut-off radius. The feature matrix of a (hypothetical) MgCa_alloy is presented as a schematic example in Figure 6.1b: the object is largely sparse; non-zero entries are represented by filled spheres, which, by the definition of the PRDF, can only occur in the columns corresponding to MgMg, CaCa, MgCa and CaMg element pairs.

The distance of two crystals is then defined as the distance induced by the Frobenius norm between such matrices

$$d(\boldsymbol{m}_i, \boldsymbol{m}_j) := |X(\boldsymbol{m}_i) - X(\boldsymbol{m}_j)|$$

and may enter a RBF kernel. In this manner, we have defined a novel global descriptor as well as a similarity measure for crystals which is invariant under translation, rotation and the choice of the unit cell. Furthermore, the periodic elements enter our representation merely as dimensions, avoiding any bias by assumed chemical similarity, which would happen in representations derived from Coulomb matrices (section 6.2).

6.4. Summary

In this section, we presented a method to represent crystal structures as vectors in a Hilbert space. Such a representation is essential for the application of machine learning kernel methods to the prediction of electronic properties, as it defines the *input space* such methods rely on. An assessment of the quality of such predictions is presented in chapter 9.

This chapter closes our high-throughput experiment design block and proceed to a description of the library of materials our HTM experiments are performed on.

7. Library of Materials

In the previous two chapters, the design of simple computational experiments to be performed on individual materials during our search for superconductors by high-throughput methods (HTMs) has been established: in chapter 5, the descriptors of superconductivity were introduced, which allow for an estimate of the superconducting properties at the moderate computational cost of Kohn-Sham DFT calcualations; in chapter 6, a path has been laid to *predict* such quantities via machine learning techniques, at *almost negligible* computational complexity.

In this chapter, we present the library providing the materials to our search, which is the second ingredient to any HTM, as the automatized experiments evidently need to be performed on a set of substances.

Moreover, we present methods to extract a subset of materials actually *usable* within our calculations in section 7.2: as is the case with every library, not all information is applicable to every possible context. A statistical analysis on this subset is presented in section 7.3.

7.1. Description of crystal structures within ICSD

Crystallographic data for our high-throughput search is taken from the comprehensive *Inorganic Crystal Structure Database* (ICSD) [51]. Entries correspond to descriptions of crystal structures reported within scientific publications; FIZ Karlsruhe, the organization responsible for the database, integrates both newly reported and legacy crystal structures on a regular bases. Both experimental results, such as structures determined via powder and single crystal diffraction measurements, and theoretically predicted materials are contained in ICSD, adjoined by a reference to the original article.

The version of the database employed for this work contains 135.468 crystal structures in total, each of them uniquely identified by a *collection code* (`coll_code`) that does not change between different versions of ICSD, and is therefore commonly used when reporting scientific results based on materials found in ICSD.

As a *relational database*, ICSD consists of a set of tables; entries in one table may reference one or more entries of another table. The central table `icsd` contains global information about each material, such as the spacegroup, the Pearson symbol (which combines Bravais lattice type and number of atoms per conventional cell), the lattice parameters and a reference to the original publication. Other tables provide information such as chemical composition, and site information (occupied Wyckoff positions including multiplicities and the associated chemical element).

While far more information is contained in the database, also used in the present work,

the aforementioned entries are the most relevant for the analysis presented in section 7.2 and section 7.3.

7.2. Criteria for excluding materials from our search

ICSD contains a huge number of materials (about 140.000), many of which are inaccessible or undesirable in our high-throughput search for superconductors. In this section, we describe reasons for excluding materials, while in section 7.3 statistical data about the remaining and therefore *usable* materials is presented.

7.2.1. Incomplete crystal structures

Limitations of the experimental measurements of crystal structures may lead to an inability to specify the exact position of all constituent elements (most prominently Hydrogen in the context of x-ray diffraction) in the primary literature.

Nevertheless, ICSD includes a significant amount of such structures, where the position of the respective atom is marked as `null`, i.e. unspecified. While in principle a position could be determined by theoretical methods involving simulations of structural relaxation, the computational demand would exceed the context of the present work. Therefore, we explicitely exclude such materials from our high-throughput search for superconductors, removing 15.224 crystals from the material set we classify as 'usable'.

7.2.2. Alloys

A second class of materials we do explicitely exclude from our 'usable' set of materials are alloys. Site occupation numbers in this case are fractional, and correspond to a *disordered* population of the given position by two or more elements. Within solid-state Kohn-Sham DFT, such materials could be in principle simulated either by constructing very large supercells or by applying the virtual crystal approximation (VCA), where pseudopotentials for such fractionally occupied sites are carefully constructed in order to represent the statistical mixture. Both approaches are prohibitively expensive in the context of our high-throughput search for superconductors, the former due to the computational demands, the latter due to the required amount of careful manual interaction and the fundamental limits of VCA. Due to this fact, we explicity exclude ICSD's 51.772 alloys from the set we classify as 'usable'.

7.2.3. Filtering by constituent elements

As ICSD contains a comprehensive set of materials, it also includes compounds of questionable practical usefullness as potential superconductors: compounds containing transuranium elements. Due to radioactive decay of such elements, they both pose a danger for researchers handling them during crystal synthesis, and would quickly introduce defects in the respective crystal structures due to the potentially low half lifes of

even their most stable isotopes. Therefore, we explicitely exclude any such compound from the ICSD's subset we consider usable.

A more technical reason for the exclusion of materials containing certain elements lies in the nature of our plain-wave-basis LDA Kohn-Sham calculations: f transition metals are particularly hard to describe by pseudopotentials reliably, therefore few have been published. Due to this fact, we explicitely exclude materials containing the lanthanides praseodymium, europium, terbium, dysprosium, holmium and erbium and the actinide thorium from our search.

Eventually, excluding both classes of compounds reduces the 'usable' subset of ICSD by another 9.024.

7.2.4. Duplicate materials

During our analysis of the crystal structures, we made the observation that ICSD also contains crystal structures that are sometimes truly identical, sometimes very close to each other, while the latter may actually be the result of measurements under pressure (a fact that cannot be easily determined from the data provided). On the first view, the presence of duplicates would merely lead to an increased computational cost due to the redundant numerical calculations, which would be desirable to avoid. However, in the context of statistical methods, including both the basic statistic analysis presented in this chapter and chapters 8/10 and the techniques employed in our (statistical) machine learning experiments (Chapter 9), such redundancy poses a serious problem due to the introduced statistical bias.

A simple graph-theory based clustering approach has been applied in order to solve this problem, due to the fact that the problem of eleminating duplicates from the dataset can be re-interpreted as a *clustering* problem: our approach identifies groups of mutually very similar materials (clusters), and represents each cluster by only one member, discarding the remaining members as they are duplicates.

Let us start our description by the criteria we apply when comparing two crystal structures C_i and C_j: both materials are counted as 'approximately equal' if

- chemical composition of the unit cells is identical
- Space groups are identical
- Label/multiplicities of the Wyckoff positions occupied by all atoms are identical
- in the case of occupied Wyckoff positions with variable components, all absolute differences lie below 0.001
- unit cell volumes differ by less than 10%, a threshold employed to separate structures under pressure.

In our graph theoretical clustering approach, each crystal structure is represented as a *vertex* N_i in a graph $\mathcal{G}$. An *edge* between two vertices N_i and N_j is established if both

Data: $\mathcal{A} = \{\boldsymbol{A}_1 \dots \boldsymbol{A}_n\}$; // Binary row vectors of adjacency matrix
Result: $\mathcal{M}$; // Totally connected subgraphs
1 **while** size($\mathcal{A}$) > 0 **do**
2 $c = \min_j |\boldsymbol{A}_j|$; // select shortest row vector
3 $\mathcal{M} = \mathcal{M} \cup \{\boldsymbol{A}_c\}$; // add it to the result set
4 **foreach** $\boldsymbol{A}_l \in \mathcal{A}$ **do**
5 $\boldsymbol{A}_l = \boldsymbol{A}_l - \boldsymbol{A}_l \& \boldsymbol{A}_c$; // Remove edges from $\boldsymbol{A}_l$
6 **end**
7 $\mathcal{A} = \{\boldsymbol{A}_i\}$ where $|\boldsymbol{A}_i| > 0$; // Remove nullvectors from set
8 **end**

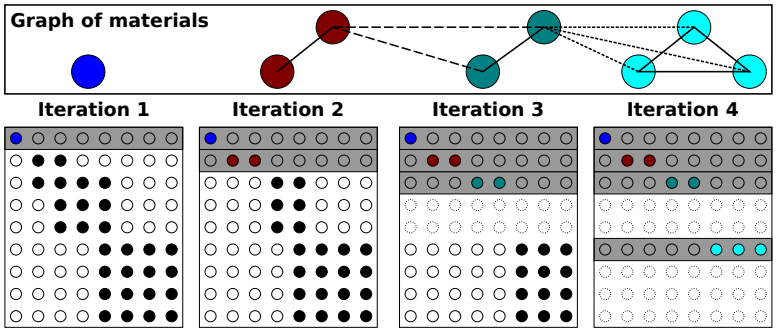


Figure 7.1.: Duplicate material detection via cluster analysis: **Top panel:** Gram-Schidt inspired orthogonalization algorithm for partitioning the graph of approximately equal (see text) materials into clusters of mutually equal materials. **Middle panel:** Example graph of materials: each vertex represents a crystal structure, which are connected if the corresponding material pair is approximately equal. Dashed lines are edges removed during the progress of the algorithm. **Lower panel:** Adjacency matrix $\mathcal{A}$ at the end of the 'While' loop (line 7 of the algorithm), which eventually partitions the example graph into 4 clusters. Members of $\mathcal{M}$ are indicated by gray background, while dotted circles indicate the nullvectors removed in line 7.

materials are considered approximately equal (an example is presented in the middle panel of Figure 7.1). $\mathcal{G}$ is then split into *cliques* (subgraphs where all vertices are directly connected to each other) $\mathcal{M}_I$, each representing a set of mutually (approximately) identical materials. Representing $\mathcal{G}$ by its ajacency matrix

$$\mathcal{A} \quad | \, a_{ij} = \begin{cases} 1 & \text{if } C_i = C_j \\ 0 & \text{if } C_i \neq C_j \end{cases} \quad ,$$

a custom orthogonalization algorithm (Figure 7.1, top panel), derived from the conventional Gram-Schmidt process (mapping the projector to a binary AND operation), is employed to split $\mathcal{G}$ into $\{\mathcal{M}_I\}$ (lower panel of Figure 7.1 displays the progress of the iterative algorithm applied to the example graph). In the final step, a representative material $m_I \in \mathcal{M}_I$ is chosen for each cluster $\mathcal{M}_I$ by selecting the member with a cell

Figure 7.2.: Classification of the 135.468 materials contained in ICSD due to the filters described in section 7.2

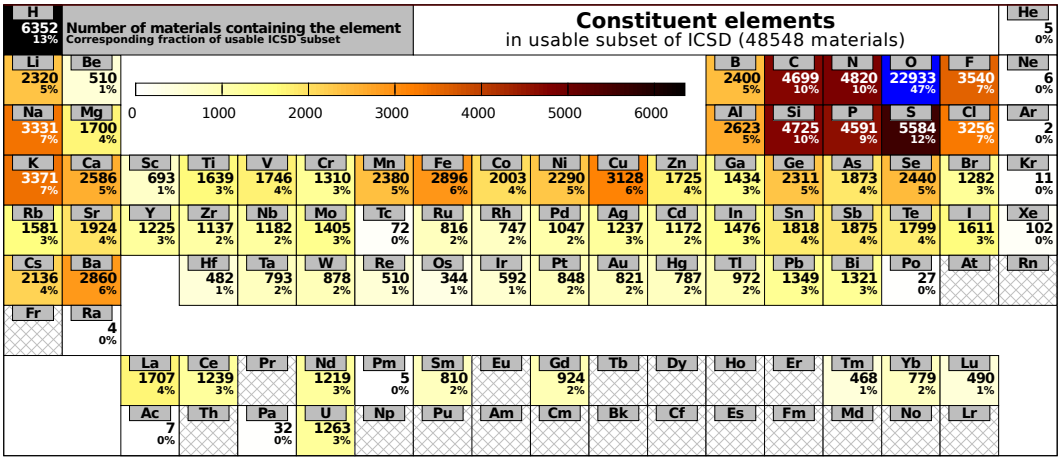


Figure 7.3.: Constituent elements of usable ICSD materials. Background color for each element corresponds to the number of materials it is found in (if the value lies outside the colormap displayed on the bottom, background color is switched to blue).

volume closest to the mean volume of all members of $\mathcal{M}_I$; if more than one material fulfills this criterium, a random choice is used to to break the symmetry.

Application of the described method identified 10.835 crystal structures which are duplicates of the remaining, not previously excluded ones.

7.3. Usable Materials: a statistical description

All 48.548 materials not excluded by the filtering criteria presented in section 7.2 are candidates in our high-throughput seach for superconductors (Fig. 7.2 summarizes the final partitioning of ICSD due to the filters). We will now present some statistical observations made on this set of candidates.

7.3.1. Chemical composition

Figure 7.3 displays a simplified overview of the chemical compositions in the 'usable' subset of ICSD. For each element, the number of materials in which it is contained is displayed, which also determines the background color. As a supplementary information, the fraction in relation to the dataset size is provided. Oxygen is found in 47% of the materials, a frequency much larger than the one of any other element, which is why it is

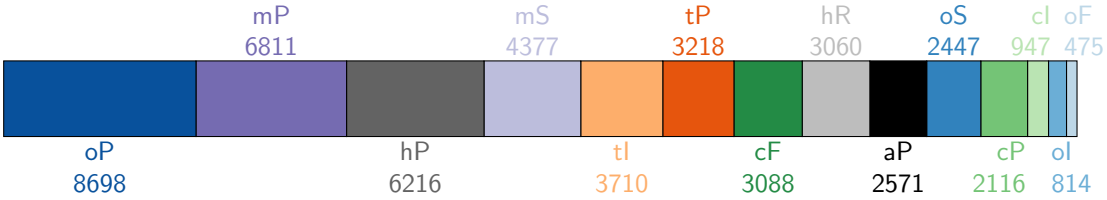


Figure 7.4.: Distribution of materials to the different Bravais lattices

assigned a color outside the regular color scale. This finding is somewhat disappointing in the context of a high-throughput search for superconductors, as there are no known conventional superconductors which contain oxygen.

Most other elements are contained in more than 400 materials each; the exceptions are noble gases, which is expected, due to their chemical inertia and non-transuranium elements without stable isotopes, namely technetium, promethium, polonium, actinium and protactinium.

The second most frequently found element is hydrogen, which, from experience, tends to be found in crystals with a molecular character (cf. section 5.4).

7.3.2. Bravais lattices

Statistical data on the Bravais lattices formed by the usable subset of ICSD is presented in figure 7.4. Materials with simple othorhombic (oP), monoclinic (mP), tetragonal (tP), cubic (cP) and base-centered monoclinic (mS) and orhorhombic Bravais lattices account for more than 57% of the available materials. This aspect of the distribution is discussed a a bit more deeply in the following section.

7.3.3. Primitive cell sizes

Neglecting the number of valence electrons in the actual chemical composition, the size (number of atoms N_{at}) of the primitive cell stands in direct relation to the number of electrons and therefore Kohn-Sham states considered while solving (1.15), which in turn strongly influences the computational demand of the simulation.

Moreover, lattice vibrational properties and coupling between electrons and phonons need to be computed for potential superconductors selected by the method described in chapter 5. As the number of phonon modes in a given system is $3 \cdot N_{\mathrm{at}}$, and essentially one linear-response calculation (section 1.2) needs to be performed for each mode, the increase of computational complexity with N_{at} even more strongly pronounced, albeit of smaller concern, due to the fact that the set of candidate superconductors will be small.

Altogether, we strongly prefer smaller systems in our high-throughput search mainly out of computational reasons.

Motivated by this fact, we investigate the distribution of usable ICSD materials among primitive cell sizes. The distribution is represented in the form of a logarithmically scaled histogram in Fig. 7.5. Three different, exponentially decaying trends dominate

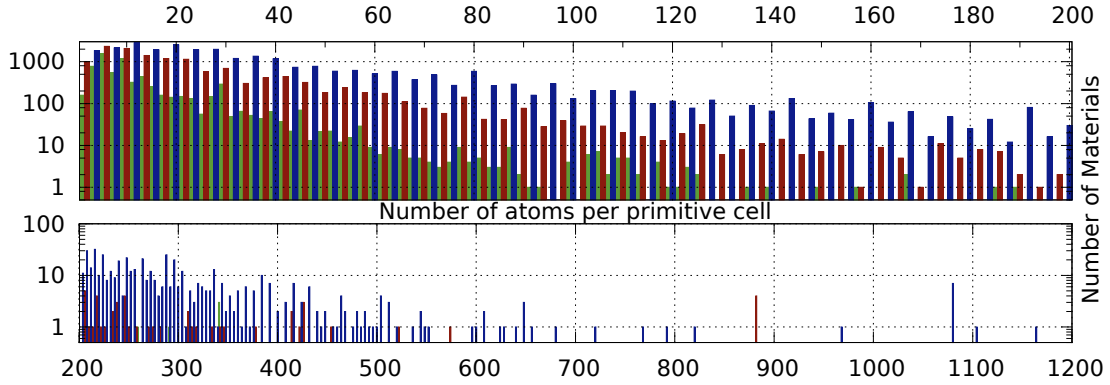


Figure 7.5.: Histogram of number of atoms per periodic unit cell in the usable subset of ICSD. Green bars correspond to odd cell sizes, while blue bars denote even numbers which are divisible by 4. Bars corresponding to the remaining even cell sizes are displayed in red.

the distribution: systems with an odd number of atoms per primitive cell, those with a cell size dividible by four, and the remaining even cell sizes. In figure 7.5, these three sub-distributions are highlighted by the application of different colors to the corresponding bars.

The trend of the number of materials exponentially decaying with the size of the primitive cell can be related to two facts. First of all, ICSD consists of data extracted from scientific publications, which already introduces a certain bias, as structurally simple crystals with a small unit cell size are more likely to be published, mainly due to experimental limitations (techniques such as x-ray crystallography become less well-resolved in the case of larger unit cells). Secondly, this trend agrees with the idea that, given a chemical composition, crystal structures with a higher symmetry are preferably formed (cf. Pauling's rules for ionic crystals [132, 133]) implying a lower number of atoms per primitive cell.

This different trends for the unit cell size moduli is surprising at first, as it would have been expected only for the sizes of the *conventional* cells due to the multiplicity induced by face-, body- and base-centered Bravais lattices. Moreover, the ratio between the total numbers of materials of the three classes 6975 : 27975 : 13598 is surprisingly close to the ratio between the moduli 1 : 4 : 2.

However, the different behaviour actually may give insight into the construction principle of a significant fraction of larger unit cells. The distribution of the materials belonging to each of the three classes among the different Bravais lattices is presented in Fig. 7.6. The most prominent differences in the number of materials belonging to the three classes arise from the contibutions of simple orthorhombic (oP), simple tetragonal (mP) and simple monoclinic (mP) Bravais lattices, giving rise to the idea that many of the larger cells may correspond to symmetry-broken (by the means of atomic substitution and/or lattice distortion) *centered* Bravais lattices: breaking some symmetries

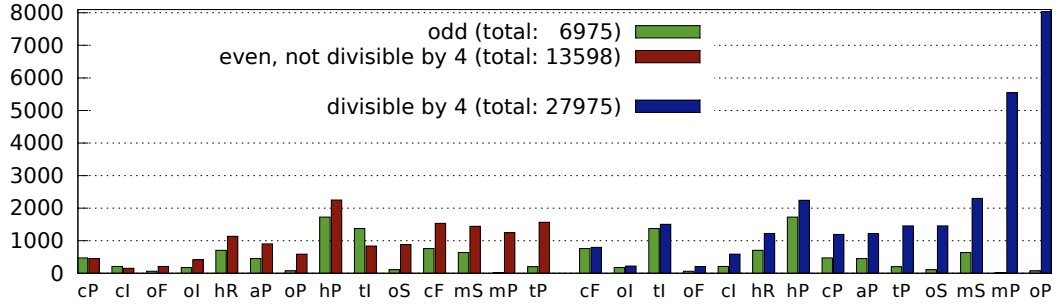


Figure 7.6.: Distribution of materials with odd, non-4-divisable and 4-divisable primitive cell sizes to the Bravais lattices (lattice labels ordered by the absolute difference to odd-primitive-cell-sized systems)

in a face-centered lattice may lead to the necessity to describe it by a supercell, corresponding to either a base-centered lattice (which introduces a factor of 2 in N_{at}) or a a simple lattice (which introduces a factor of 4 in N_{at}). A corresponding consideration for symmetry-broken supercells of base- and body-centered lattices would introduce a factor of 2, when the broken symmetry neccessitates the description by a simple Bravais lattice.

7.4. Dataset Materials

For this work, calculations have been performed for all ICSD materials classified as usable (section 7.3), with up to 8 atoms per primitive unit cell. This subset consists of 8.212 (15.457 when including duplicates) crystal structures.

For each of the structures, self-consistent Kohn-Sham calculations were performed, with the computational parameters described in subsection A.2.1. The calculations were successfull for 8.071 crystal structures, which form the *dataset* for the statistical evaluation in our high-throughput search.

In this section, we introduce these *dataset materials*, by a similar statistical description that has been performed for the usable subset of ICSD in section 7.3.

7.4.1. Successfull simulations

With the help of the Job Supervision Framework (JSF, subsection A.2.2), results could be successfully obtained for 8.071 different materials, where failure to successfully perform simulations for the remaining 141 materials were related to errors in the materials' description within ICSD. Effectively, data has been obtained for about 17% of ICSD's usable subset (section 7.3). Statistics regarding failure and automatic error recovery, which is essential to a computational high-throughput search, can be found in subsection A.2.3.

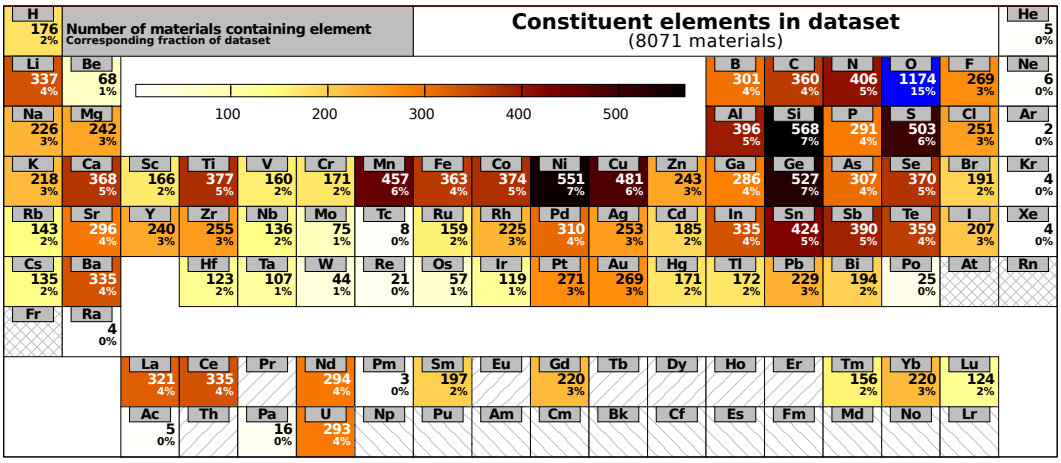


Figure 7.7.: Constituent elements within the dataset. Background color corresponds to absolute number of materials containing the respective element; if the number lies outside the colorbar displayed on the bottom, color is switched to blue

7.4.2. Statistical description of the materials

We will now present a brief statistical description of the crystal structures in the *dataset*, also to allow a better interpretation of the statistical descriptions of the properties found in chapter 8.

Chemical composition

Figure 7.7 displays a simplified overview of the chemical compositions in the dataset. For each element, the number of materials in which it is contained is displayed, which also determines the background color, and as a supplementary information, the fraction in relation to the total number of materials is provided. Comparing to Figure 7.3, oxygen is far less dominant, contained in 15% of these smaller materials, than it is in the overall ICSD, where 47% of the materials contain it. Moreover, the fraction of materials containing hydrogen, giving rise to the concern of a large number of molecular crystals in the earlier discussion about the whole ICSD, is now comparable with the other elements, as molecular solids tend to occur with larger cell sizes, necessary to describe the constituent molecules. The low representation of noble gases and radioactive elements has been discussed in subsection 7.3.1.

Summarizing, 80% of the periodic elements are present in 100 materials each, which suggests that the dataset presents a good sample of a wide range of chemical compositions.

Crystal structures

In Figure 7.8, the distribution the datasets' members among different primitive cell sizes is displayed; the trends regarding cell size moduli described in subsection 7.3.3 is also

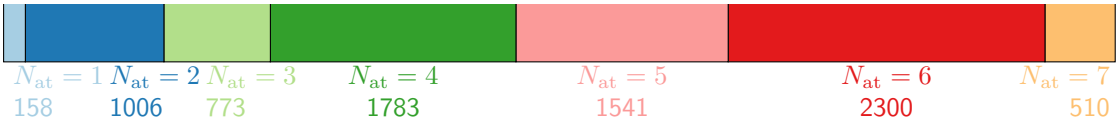


Figure 7.8.: Distribution of dataset members to unit cell sizes

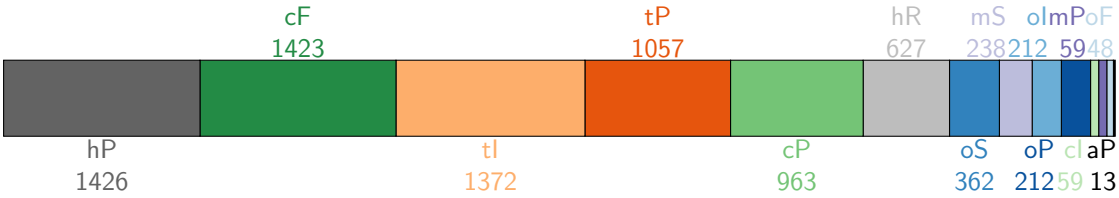


Figure 7.9.: Distribution of dataset members to bravais lattices

observable here, in the limit of smaller cells, however overlayed with the limiting factor that due to the finite number of periodic elements, combined with the low number of possible sites, only a lower number of elemental combinations could be built.

The distribution of the small systems among bravais lattices largely differs from the overall distribution within ICSD (Figure 7.4), also in agreement with the discussion in subsection 7.3.3: while the dataset spans only 16% of ICSD, it accounts for 46% of the simple and face centered cubic (cP/cF), 37% of the body-centered tetragonal (tI), $\frac{1}{3}$ of the simple tetragonal (tP) structures and about 20% of the hexagonal (hP) and rhombohedral (hR) structures.

7.5. Prediction of new materials via element substitution

The discovery of new materials is an integral part of many computational high-throughput studies, especially when materials are optimimized with respect to desired properties. While there are ab-initio methods available for the prediction of likely crystal structures with a given chemical composition [134–138] such methods are, due to the dimensionality of the problem, computationally expensive.

Before the advent of the ab-initio methods, crystal structure prediction was mainly performed by heuristic rules, such as the ones by Pauling [132] (relation between ionic radii and structure in ionic crystals), Hume-Rothery [139] (relation between valence electrons per atom and the crystal structure) and Pettifor [140] (structure of binary compound predicted by element pair coordinate determined with an appropriate mapping of the atomic numbers).

Datamining approaches to the prediction of new materials [141, 142] determine rules on the basis of statistical data obtained on large sets of existing materials. Within this section, we present such an approach.

We consider the creation of a new material from a known one by *element substitution*, i.e. the replacement of one or more atoms of type A by a different periodic element B without fundamentally changing the crystal structure. As an example, the non-superconductor AlB_2 is related to the superconductor MgB_2 by such a substitution,

accompanied by a slight adjustment of the interlayer distance.

The majority of changes in chemical composition lead, due to chemical/bonding differences, either to different crystal structures or no thermodynamically stable structures at all. However, statistics performed on known and therefore existing and stable structures found within ICSD allow us to synthesize information on probable candidates for substitutions with the desired properties.

In the first subsection, we provide our definition of element substitution relations between materials and a basic overview on such relations among materials contained in ICSD. Extracted information regarding the substitution properties among periodic elements and its application to the construction of new materials is presented in the final part of this section.

7.5.1. Definition of element substition

In this subsection, we provide a definition for the *element substitution* relations between materials used later within this section. In summary, we consider pairs of materials with identical crystal structure (defined by *structure prototype*), which differ by the substition of one periodic element by another one. Lattice lengths and angles are left as a degree of freedom, as long as the symmetry is conserved.

It is convenient for our study to start with our definition of a *structure prototype*. The concept itself is tentatively covered within most solid state physics courses, when structure types like rocksalt, cesium chloride, zincblende or wurtzite are introduced: while the names are derived from minerals with a particular chemical composition, also other constituent elements may crystallize in analogous structures.

Each such prototype is not defined by a particular combination of periodic elements, but by *bravais lattice*, *symmetry* and *positions of the basis atoms*. All this information is compactly represented by the *space group* (independent of the lattice lengths and angles) and the set of occupied *Wyckoff positions* (independent of the type of atom actually occupying them). Consequently, we define materials with identical structure prototype $T^{(I)} = T^{(I)}$ as those with identical space groups and identical sets of occupied Wyckoff positions, considering a tolerance interval ϵ on internal positions. Given its structure prototype $T^{(M)}$, any material M can be uniquely described by the means of an *occupation vector*

$$\boldsymbol{X}^{(M)} = (X_1^{(M)}, X_2^{(M)}, \dots X_N^{(M)}) \quad | X_i^{(M)} \in \text{periodic elements},$$

whose components define the atoms occupying the ordered set of N Wyckoff positions.

Within this work, we only consider substitution relations involving a single pair of periodic elements, as we base our construction principle for new materials on such relations. Furthermore, *elemental solids* of element A are only considered regarding partial substitutions, i.e. the structure prototype is required to have more than one occupied Wyckoff position and a substitution by element B is required to leave a subset of A unchanged. Note that some substitution relations between materials may not be found by this approach, as identical systems may be described by different prototypes, a fact

Figure 7.10.: Element substitution in (complete, non-alloy) ICSD materials. 44% of the non-duplicate[1] materials are interrelated by single-element substitutions $A \to B$.

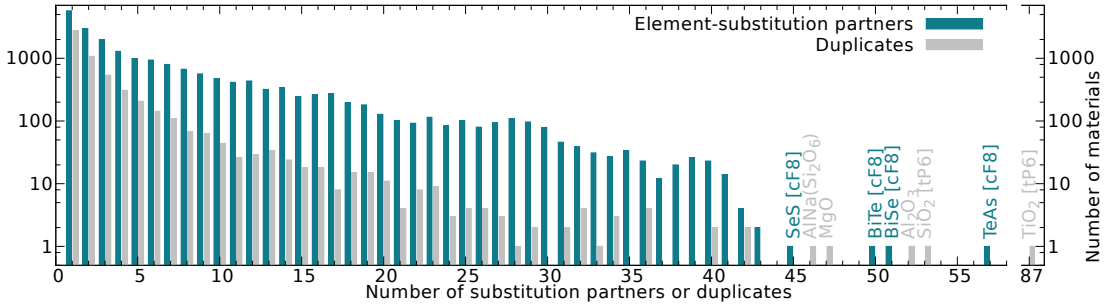


Figure 7.11.: Number of other materials related to each material by element substitution (7.2) among ICSD materials (duplicate counts have been included for completeness).

that we neglect in this first approximation. Detecting also such cases would require a far more complex approach, employing topological transformations in the comparison process.

An analysis of the previously outlined substitution relations among ICSD materials has been performed, excluding incomplete materials (subsection 7.2.1) and alloys (subsection 7.2.2). The result of this analysis is

$$S_{AB}^{IJ} = S_{BA}^{JI} := \begin{cases} 1 & \text{if } T^{(I)} = T^{(J)},\ \boldsymbol{X}^{(I)} \text{ contains } A \text{ and } \boldsymbol{X}^{(J)} = \boldsymbol{X}^{(I)} \,|\, A \to B \\ 0 & \text{otherwise} \end{cases} \tag{7.1}$$

for all materials I, J within ICSD and any pair of elements A, B. The special case $T^{(I)} = T^{(J)}$ and $\boldsymbol{X}^{(I)} = \boldsymbol{X}^{(J)}$ describes duplicate materials, and only one of all instances is kept for the analysis in order to avoid statistical bias (cf. subsection 7.2.4).

In Figure 7.10, a first overview is presented[1]: A least one other material related by substitution has been detected for 20.500 materials, corresponding to 44% of the remaining materials, which demonstrates that such a relation between different materials is a rather common phenomenon. This broad subset of materials is the basis for our element-substitution analysis and predictions. Partial substitution is found only in 600 cases, as the detection of many such cases would actually require comparison of materials belonging to different structure types, considering that many partial substitutions may break the symmetry of the original system.

[1] in comparison to subsection 7.2.4, the number of duplicates is increased, as the present analysis is independent of the cell volumina, and no materials are excluded on the basis of constituent elements.

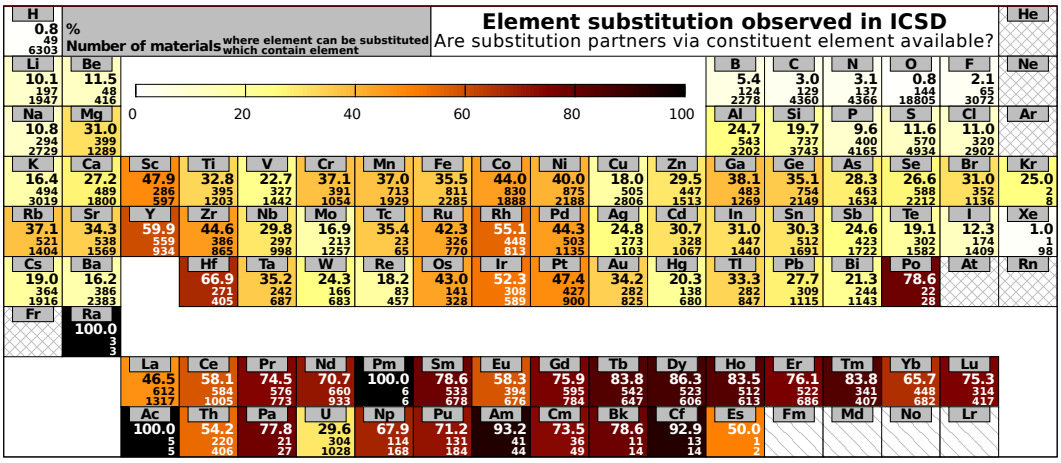


Figure 7.12.: Substitution probability R_A (7.3) for all elements in the periodic table. Numerator and denominator of (7.3) provide information on the quality of statistics for each element.

The number of other materials J related to a material I by element substitution is

$$\sum_{J \neq I} \sum_{A,B \neq A} S_{AB}^{IJ}. \tag{7.2}$$

The resulting distribution is presented in Figure 7.11. The vast majority of materials (note the logarithmic axis scaling) is found to have only few other materials related to them; less then 100 examples are found related to more than 30 other materials each. Only few materials, all mutually related by substitution and belonging to the family of binary selenides and tellurides in rocksalt structure, are related to 45 and more materials by single-element substitutions.

We have now outlined the basic definitions of our substitution analysis, and presented statistical data providing some support for its significance. In the following, we present the substitution properties on the level of the periodic elements.

7.5.2. Substitution probability of an element

A first quantity useful to predict new structures is the probability to find a thermodynamically stable material by replacing an element A at all[2]:

$$R_A := \frac{1}{N_A} \sum_I \Theta_0 \left(\sum_{J \neq I} \sum_{B \neq A} S_{AB}^{IJ} \right), \tag{7.3}$$

where N_A is the number of materials which contain A. The step function with $\Theta_0(0) := 0$ will contribute to the outer summation if there are one or more other materials related to I by substituting A by another element. R_A is therefore normalized to $0 \leq R_A \leq 1$.

[2] the question of likely substituents B will be treated separarately in subsection 7.5.4

In the limit of *all possible materials*, R_A could be interpreted as a *probability* to reach a new, stable crystal structure by a substitution of element A; however, as ICSD consists only of a subset of all thermodynamically stable materials, we are expected to be far from this limit and the numbers could be merely interpreted as trends.

In Figure 7.12, R_A is displayed for all elements of the periodic table. The highest probabilities are found for the lanthanides and actinides, with $R_A > 60\%$ for most elements; substitutions are most often observed among elements in the same series, consistent with the obvious chemical similarity [143, 144].

On the other hand, the lowest R_A for hydrogen and the first-row p elements, with a maximal $R_A = 5\%$ for B. This finding is consistent with chemical experience, that in fact speaks about the *first-row anomaly* [145, p. 245]: ,,Properties of elements in this row are frequently significantly different from properties of other elements in the same group". Among the properties referred to are atomic radii, electronegativity and also bonding behaviour [146], which in consequence mean that a substitution by elements found within other rows of the p block is far less likely to leave the crystal structure unchanged than substitutions between p block elements of the remaining periods. Among the remaining main group elements, R_A lies between 10% and 38%, while d transition metals reach between 16% and 67%.

7.5.3. Element pair example count and noise reduction

To proceed with our analysis we define an important quantity that is simply the number of examples for particular substitution $A \to B$:

$$S_{AB} := \sum_{I,J \neq I} S_{AB}^{IJ}, \tag{7.4}$$

where I and J run over all non-duplicate materials. By construction, this quantity is symmetric $S_{AB} = S_{BA}$.

However, before a further analysis, one needs to consider *noise* in the extracted data: materials reported may be subject to measurement errors, theoretical predictions never confirmed by experiment or trivial data entry errors made when including a material in the database. A simple prediction model for new materials via element substitution (subsection 7.5.4) based on noisy data will suffer a bias away from the more relevant information. We apply the most straightforward scheme for a *noise reduction* and impose a simple threshold θ on S_{AB}, such that $S_{AB} \geq \theta$ for the remaining element pairs. The choice of θ presents a tradeoff between noise reduction and loss of information for the further statistical analysis. In fact, the majority of element substitutions $A \to B$ detected by our method is only rarely observed. $S_{AB} < 5$ is found in 50% of the cases, and even $S_{AB} = 1$ in 20%. As a sidenote: there are more than 592 Si $\longleftrightarrow$ Ge examples, 474 S $\longleftrightarrow$ Se and 446 Ni $\longleftrightarrow$ Co ones; various inter-lanthanide pairs dominate the remaining region of high S_{AB}. Cross-validation considering a set of test elements suggests the conservative choice of $\theta = 3$, i.e. any pair detected only once or twice is classified as noise and removed from the set; this process shrinks the original set of pairs by 32%. Any remaining noise will be accounted for by the predictor.

7.5.4. Substitution partner probability

In subsection 7.5.2, we had defined R_A (7.3), the probability that an element A can be substituted at all. The question to be answered in the following is the prediction of substituent elements B.

As stated previously, one must consider biases within ICSD: periodic elements occur at rather different frequencies (Figure 7.3), certain structure types and compositions have received stronger researchers' attention than others. Therefore, our substitution analysis is performed in terms of the *substitution partner probability*

$$P_{AB} := \sqrt{P(B|A)P(A|B)} \tag{7.5}$$

where

$$P(B|A) \approx S_{AB}/S_A \quad \text{with} \quad S_A := \sum_{B'} S_{AB'} \tag{7.6}$$

is the conditional probability of B being the substituent if A can be substituted. Using (7.6), (7.5) assumes the simple form

$$P_{AB} \approx \sqrt{\frac{(S_{AB})^2}{S_A S_B}}. \tag{7.7}$$

As it is defined by a product of probabilities, $0 \leq P_{AB} \leq 1$, where the upper bound is reached in case A and B substitute each other *exclusively*. The normalization with respect to S_A and S_B has an important effect: statistical bias due to the large difference in how often the periodic elements A and B appear within ICSD materials (Figure 7.3) is compensated; the noise-cancelling threshold θ applied in S_{AB} (subsection 7.5.3) prevents, on the other hand, an overemphasis of noise and insufficiently sampled elements. A compact representation of P_{AB} is displayed in Figure 7.13, using the atomic number of A and B as coordinates. The matrix P_{AB} exhibits an irregular-block structure, with blocks corresponding to periods in the periodic table. The submatrices around the main diagonal correspond to intra-row substitutions, while the ones along the secondary diagonals contain inter-row substitutions.

Interpreting P_{AB} as a measure of *chemical simarilarity*, the statistics presented in Figure 7.13 are directly related to Mendeleev's[147] and his successor's considerations when constructing the periodic table of elements: when ordering the elements by atomic mass (or as we know today: nuclear charge Z), periodic behaviour of chemical properties is observable; one of the fundamental considerations were the molecules and compounds formed by an element. Periodic behaviour is the most evident in the strongly enhanced substition correlations along inter-row diagonals of the s and p blocks.

When focusing on the intra-row blocks, also the special properties of transition metals, be it d or f block ones, can be observed, by the (in the comparison to intra-s or intra-p) pronounced intra-d/f P_{AB}. This is the most prominent in the case of lanthanide-lanthanide or actinide-actinide substitions (with the exception of the statistically underrepresented Promethium), which are far more likely to occur than substitutions by any non-lanthanide/actinide element; second to this are the intra-3d substitutions.

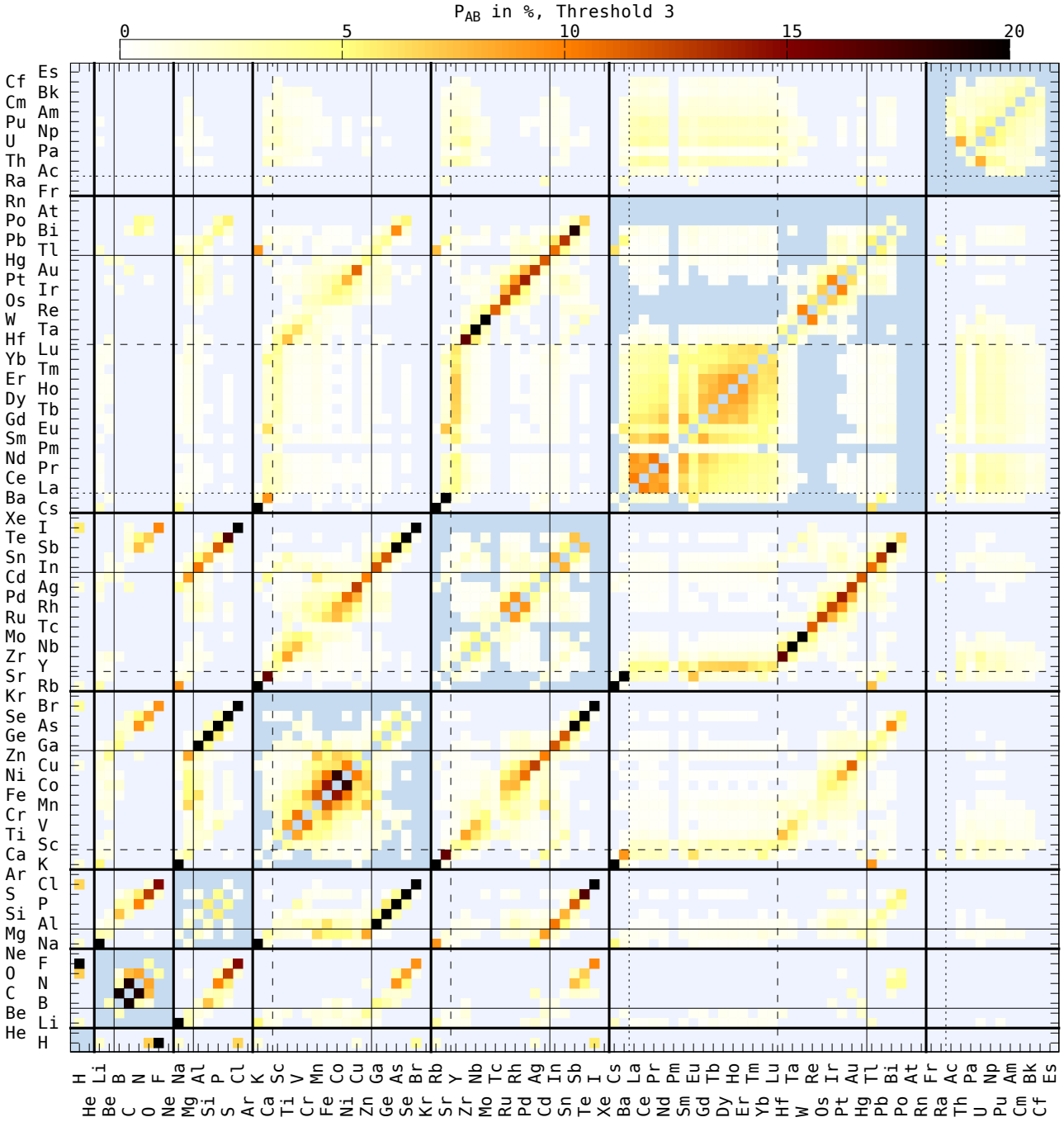


Figure 7.13.: Substitution partner probability P_{AB} (7.7) x/y Coordinates correspond to (A, B) element pairs in atomic number order; lines mark period and block boundaries. The line style indicates subshell of following block: thick solid: s (and marks therefore periods), thin solid: p, dashed: d, dotted: f. Color is defined by P_{AB}, with threshold $\theta = 3$ (subsection 7.5.3) for noise reduction.

Substitutions of Lanthanides by non-f elements B are the most often observed if B is either a member of the 3rd or alkaline earth groups, consistent with the lanthanides position between s and d blocks of the periodic table.

Substitution correlation of specific elements

So far, agreement of our data with basic chemical knowledge has been established. However, substitutions beyond this are of special interest in the construction of truly new materials, as that basic knowledge is wide-spread, and has been applied before in the construction of at least some of the materials serving as input to our study. In the following, we will discuss the substition properties of a few selected elements (Figure 7.14).

Hydrogen is the lightest of all elements. Our statistics (top left panel) shows H most often substitutes for the halogens F,Cl, Br, I and O and therefore serves as an electron acceptor. Only second to these, the alkalines Li,Na,K and, which is a bit surprising, the late-d transition metals Cu and Ag are observed as substituents, where H could be qualified as an electron donor. However, one must remember that the overall probability (7.3) of H substitutions is one of the lowest observed for any periodic element (cf. Figure 7.12).

Sodium is chosen as a representative for the alkali metals group (top right panel). The highest P_{AB} (7.7) are observed for alkali metal substituents in the neighbouring periods, a result expected by basic chemical knowledge. Substition correlation for Ag lies in the same range as for Cs, which can be related to the Na−Ag similarity in covalent radii and s^1 atomic valence configuration, in the case of silver above a closed d subshell. For Tl and Ca, g_{AB} is approximately two orders of magnitude lower than for the alkali neighbours, and only spurious contributions are observed among other metals. The other alkali metals show very similar behaviour, and one can conclude that substituting any alkali by a non-alkali will in most cases either fail or introduce fundamental changes to the crystal structure.

Magnesium, as a representative of the alkaline earth metals (center left panel), features a much broader set of substituents. The ones with the highest P_{AB} (7.7) are cadmium and zink, sharing the s^2 valence (in both cases above a closed d subshell); in the case of Zn, covalent radii agree, while in the case of Cd, there is similarity in metallic radii. Manganese and later $3d$ transition metals are next in g_{AB}, only then follow the alkaline earths of the neighbouring periods. Almost one order of magnitude less is observed for the post-transition metals Al and In; d transition metals Hg and Ag are the remaining substituents with significant contributions. Only spurious substitutions are observed for other d or f transition metals and group 13 post-transition elements.

Boron represents an example for the p block elements (center right panel), albeit an untypical one due to the first-row anomaly. Strongest g_{AB} is observed for carbon, which would correspond to approximately 20% mutual substitution $\sqrt{g_{AB}}$; p block members Si, Ga, P and Ge follow. The remaining elements with significant correlation are the adjacent alcaline earth Be, other members of the p block follow. Substitution by d transition metals are spurious at best. While not as extreme as in the case of H, the probability to substitute B *at all* is far lower than for the majority of other elements.

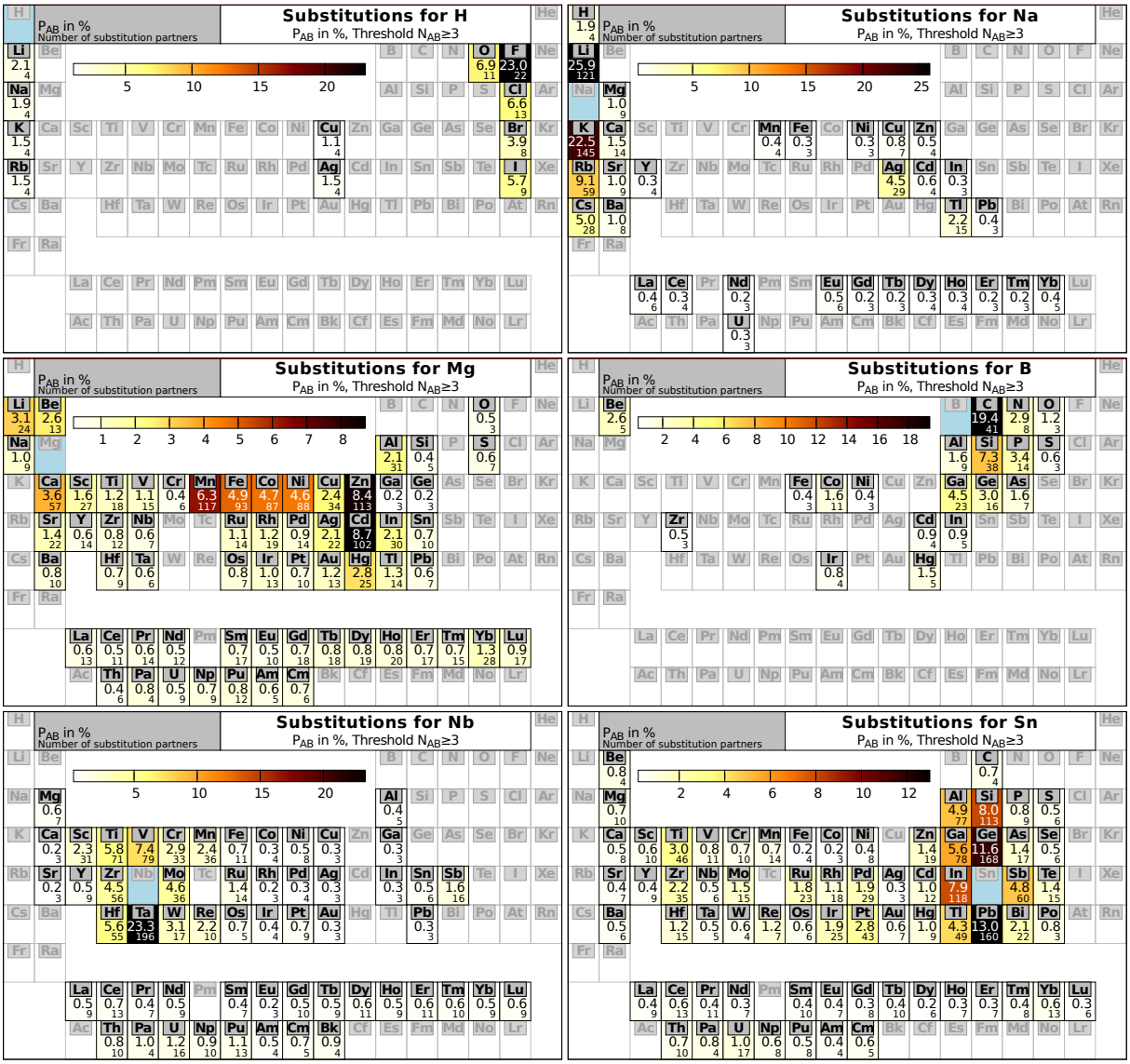


Figure 7.14.: Substitution partner probability P_{AB} (7.7) (color and first number) and S_{AB} (7.4) (second number) for $A \in \{\mathsf{H},\mathsf{Na},\mathsf{Mg},\mathsf{B},\mathsf{Nb},\mathsf{Sn}\}$; dimension B is layed out as a periodic table.

Niobium is taken as a representative of the d transition metals (bottom left panel). P_{AB} (7.7) is highest for Ta and V, the other group 5 elements. The remaining dominant substituents are found in the adjacent subgroups of the periodic table, namely Ti, Hf, Zr (group 4) and Mo, W, Cr (group 6), followed by the second-next subgroup elements Mn, Re and Sc. This sequence of substitutions is in general agreement with present knowledge about transition metal chemistry. However, the occurence of the metalloid antimony (Sb) as a significant substituent for Nb is surprising at first, but can be explained by the similarity in both atomic and covalent radii, despite the rather different valence configurations.

Tin serves as an example for a post-transition metal (bottom right panel), member of group 14 of the periodic table. P_{AB} (7.7) is the highest in other group-14 post transition elements, namely Pb, Ge, and Si, followed by group-13 post-transition metals In, Ga, Al, Tl and the metalloid Sb. g_{AB} is almost an order of magnitude lower for a small subset of the d metals, as well as the remaining post-transition metals and metalloids (with the exception of Po), without an obvious connection to the atomic/ionic radii or electronegativity.

7.5.5. Conclusion

In this section, a method for the construction of new materials, based on datamining performed on the inorganic crystal structure database (ICSD), was presented. The method is based on the assumption that the crystal structure can be invariant (apart from an adjustment of the lattice parameters) under single-element substitutions.

Statistical properties of such structure-invariant substitution relations were estimated with the help of a large set of existing materials for each pair of elements. A global overview on the statistical data was presented, among others showing that in principle the periodic table of elements could be reconstructed from it, in a similar manner as in Mendeleev's original work (solely based on statistics, and before the discovery of quantum mechanics). The overview was followed by a more in-depth review of substitions for a small, but representative subset of selected elements.

The central quantities for the prediction of new materials from existing ones by element substitution are the probability to replace an element R_A (7.3), and the substitution partner probability P_{AB} (7.7) giving the most likely substituents B for A.

In the final subsection, a method to employ the statistical data for the discovery of new materials was proposed.

7.6. A modified Pettifor chemical scale from data mining

7.6.1. Introduction

The organization of the chemical elements in a "table" has fascinated and motivated scientists for the best part of two centuries. The traditional representation of the periodic table has a two-dimensional structure, with elements arranged in periods and groups. This arrangement not only puts into evidence the chemical similarity between atoms, but

also reflects the basic quantum-mechanical character that rules atomic physics. Since the seminal works of Lothar Meyer and Dimitri Mendeleev, hundreds of such two- (or even higher-) dimensional representations have been put forward, featuring spirals, circles, cubes, etc. Moreover, one can arrange the elements according to their atomic properties, or choose to put in evidence other molecular or solid-state properties.

It is true that the best description of the relationship between the chemical elements requires two (or more) dimensions. However, in many practical cases, one requires a much simpler, one-dimensional ordering where elements that are chemically similar occupy neighboring positions. In this section we are concerned with one such ordering, already introduced by Pettifor 30 years ago [148], and that is today extensively used in the modern field of accelerated materials design and high-throughput calculations (see, e.g., [101, 149, 150]).

Pettifor's original interest was on the structural stability of binary AB compounds [148]. Binary compounds crystallize in 34 different structure types. If we assign a different symbol to each structure type and plot it for each A and B we obtain a so-called structure map [151]. The problem that Pettifor tried to solved was how to order the chemical elements in order to achieve the best structural separation within such two-dimensional plot. There had been several previous attempts to achieve such separation using properties like the core radius, the electronegativity, the number of valence electrons, etc. Unfortunately these approaches not only led to high-dimensional structure maps (difficult to plot and visualize), but also to a rather disappointing structural separation. Pettifor's solution was rather elegant, but also quite radical. He neglected all theoretical considerations and constructed a fully phenomenological one-dimensional ordering of the elements that provided a near-perfect structural separation of the AB binaries. Further work showed that this was also true for other binary A_xB_y systems [140].

Pettifor had at his disposal 574 binary AB compounds plus a few hundred other binaries phases. Today we have available the experimental crystal structures for at least two orders of magnitude more compounds. This information can be found in several databases such as the inorganic crystal structure database (ICSD) database [52], or the crystallography open database [152]. In this section we show that we can use Pettifor's idea that chemical similarity manifests itself in the formation of similar structures and use the wealth of new information to improve its original scale.

7.6.2. A mathematical definition of the (modified) Pettifor scale

Let us remember that Pettifor constructed his scale by trying to separate the different crystal structures of binary compounds AB in a binary diagram [148]. Besides the knowledge of the crystal structures of AB compounds, his main tools were his formidable chemical intuition and trial and error. In our work we give a step beyond, and use the statistical analysis of section 7.5 to perform the task of creating a chemical scale. This has the advantage of being completely unbiased with respect to possible (human) prejudices, and of assuring the optimal (or at least a very good) ordering based on the totality of the available data.

Having obtained the matrix describing the probability of a successful substitution of

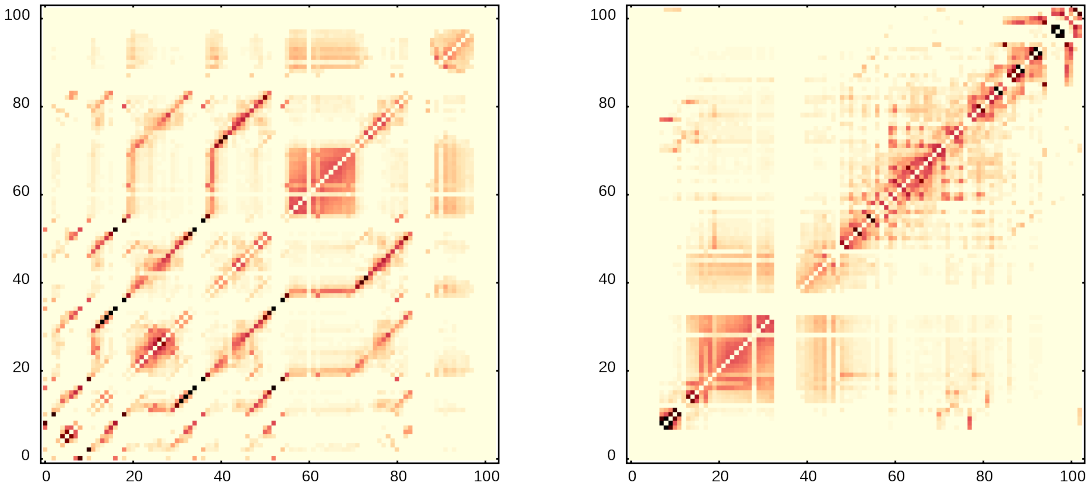


Figure 7.15.: The probability matrix P_{AB} using the atomic number (left) and the Pettifor scale (right) to order the chemical elements. The color range goes from white (zero entry) to black (entry $>= 0.3$). The maximum entry is ≈ 0.62 for the pair Br–Cl.

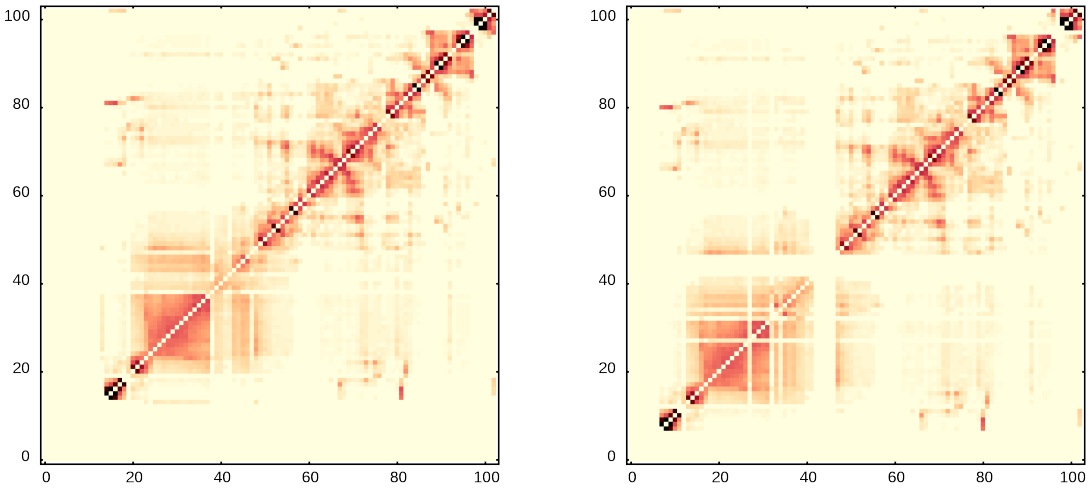


Figure 7.16.: The probability matrix P_{AB} using our chemical scale optimized with genetic algorithms (left) and our proposed modified Pettifor scale (right) to order the chemical elements.

an element A by an element B, we can now proceed to the construction of a new Pettifor map. The idea is very simple: If two chemical elements are similar, then it is probable that one can substitute one by another in a given crystal structure. This will lead to a large entry in the matrix element (A, B). If we can order the chemical elements such that the large matrix elements are close to the diagonal, this implies that similar elements will occupy neighboring positions in the chemical scale.

The left panel of Figure 7.15 shows the matrix P_{AB} using as ordering of the chemical

elements the atomic number. The first striking evidence is that the matrix has a very geometrical structure and it contains very large entries far away from the diagonal (cf. subsection 7.5.4).

Clearly, using the Pettifor scale (see Table 7.1) to order the chemical elements yields a matrix that is in a much more diagonal form (see right panel of Figure 7.15). Now, not only the lanthanides form a clear structure, but several other groups are also clearly visible across the figure. It turns out that the Pettifor scale is already a very good solution to our ordering problem (we will show actual quantitative evidence below).

Finding a numerical framework to make the matrix more diagonal is very simple as soon as we recognize that our problem is similar to the reduction of the bandwidth of a sparse matrix. As this plays a very important role in the solution of large linear systems, it has been studied intensively since the original Cuthill-McKee algorithm in 1969 [153]. This problem is also related to the famous traveling salesman problem. In fact, the traveling salesman has to find a path through a certain number of cities (i.e., an ordering of the cities) that minimizes the total travel distance. We have to find a path through the chemical element space that maximizes the diagonal character of the matrix.

The traveling salesman problem is a hard problem (NP-complete), and the time to find the optimal solution grows exponentially with the number of cities. However, many strategies have appeared over the years to obtain good solutions. We decided to use genetic algorithms, mainly due to their simplicity. The first step in using genetic algorithms is defining the objective function to be optimized. Following several numerical experiments we selected the following function:

$$\mathcal{F} = - \sum_{A,B \neq A} \frac{P_{AB}}{|i_A - i_B|} \,, \tag{7.8}$$

where P_{AB} is defined by Equation 7.7, i_A is the position of element A in the ordering, and the sum runs through all pairs such that $A \neq B$. This choice gives increased weight for entries close to the diagonal, while not penalizing too much small entries far from the diagonal. Obviously, the function $\mathcal{F}$ has to be minimized.

The second step is to define a gene. We take simply a list of 103 entries with the natural numbers from 1 to 103, indicating the order in which the chemical elements should be arranged. For the crossover operator we first select randomly a segment of one of the parents that is passed to the same position to the child, and then fill the voids using the gene of the second parent by removing the entries already contained in the child gene. We tried as mutation operations: (i) swapping two random elements in the gene or (ii) moving an element from one random position in the gene to another. We found that the second choice was greatly superior in our simulations. The mutation rate was set to 20%. We note that, as our matrix has a relatively low dimensionality, we did not have to use more sophisticated and efficient genetic algorithms such as the ones from Ref. [154].

To solve the problem of the elements for which no information exists in ICSD we moved them all to the beginning of our gene. Furthermore, and in order to have an

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Z	H	He	Li	Be	B	C	N	O	F	Ne	Na	Mg	Al	Si	P	S	Cl	Ar	K	Ca	Sc	Ti	V	Cr	Mn	Fe
P	He	Ne	Ar	Kr	Xe	Rn	Fr	Cs	Rb	K	Na	Li	Ra	Ba	Sr	Ca	Yb	Eu	Y	Sc	Lu	Tm	Er	Ho	Dy	Tb
GA	He	Ne	Ar	At	Rn	Fr	Es	Fm	Md	No	Lr	Kr	Xe	Pm	Cs	Rb	K	Na	Li	Ra	Ba	Sr	Ca	Eu	Yb	Lu
$\mathbf{P_m}$	He	Ne	Ar	Kr	Xe	Rn	Fr	Cs	Rb	K	Na	Li	Ra	Ba	Sr	Ca	Eu	Yb	Lu	Tm	Y	Er	Ho	Dy	Tb	Gd

n	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
Z	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr	Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te
P	Gd	Sm	Pm	Nd	Pr	Ce	La	Lr	No	Md	Fm	Es	Cf	Bk	Cm	Am	Pu	Np	U	Pa	Th	Ac	Zr	Hf	Ti	Nb
GA	Tm	Y	Er	Ho	Dy	Tb	Gd	Sm	Nd	Pr	Ce	La	Ac	Am	Cm	Bk	Cf	Pu	Np	U	Th	Pa	Sc	Zr	Hf	Ti
$\mathbf{P_m}$	Sm	Pm	Nd	Pr	Ce	La	Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr	Sc	Zr	Hf	Ti	Ta

n	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78
Z	I	Xe	Cs	Ba	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	Hf	Ta	W	Re	Os	Ir	Pt
P	Ta	V	Mo	W	Cr	Tc	Re	Mn	Fe	Os	Ru	Co	Ir	Rh	Ni	Pt	Pd	Au	Ag	Cu	Mg	Hg	Cd	Zn	Be	Tl
GA	Nb	Ta	V	Cr	Mo	W	Re	Tc	Os	Ru	Ir	Rh	Pt	Pd	Au	Ag	Cu	Ni	Co	Fe	Mn	Mg	Zn	Cd	Hg	Be
$\mathbf{P_m}$	Nb	V	Cr	Mo	W	Re	Tc	Os	Ru	Ir	Rh	Pt	Pd	Au	Ag	Cu	Ni	Co	Fe	Mn	Mg	Zn	Cd	Hg	Be	Al

n	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	
Z	Au	Hg	Tl	Pb	Bi	Po	At	Rn	Fr	Ra	Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr	
P	In	Al	Ga	Pb	Sn	Ge	Si	B	Bi	Sb	As	P	Po	Te	Se	S	C	At	I	Br	Cl	N	O	F	H	
GA	Al	Ga	In	Tl	Pb	Sn	Ge	Si	B	C	N	P	As	Sb	Bi	Po	Te	Se	S	O	I	Br	Cl	F	H	
$\mathbf{P_m}$	Ga	In	Tl	Pb	Sn	Ge	Si	B	C	N	P	As	Sb	Bi	Po	Te	Se	S	O	At	I	Br	Cl	F	H	

Table 7.1.: The different chemical scales: **Z** is the atomic number; **P** is the Pettifor scale; **GA** is the scale stemming from the genetic algorithms; $\mathbf{P_m}$ is the modified Pettifor scale proposed in this work.

easier comparison with the Pettifor scale we decided to fix the two end-points of our scale to be Kr (the first rare-gas for which we have data) and H (number 103 in the Pettifor scale). We checked that this arbitrary choice does not have a significant impact in the value of the minimum objective function.

7.6.3. Results

We run a series of simulations using a pool of 200 genes in our population that were evolved for around 500 generations. We used as starting points random genes, the Pettifor scale, and the ordering by atomic number. Our best result is shown in the left panel of Figure 7.16 and in Table 7.1.

We can use the numerical value of Equation 7.8 in order to have a quantitative assessment of the quality of our chemical scale. A random ordering of the chemical elements leads to a value of $\mathcal{F}$ typically between -3 and -4. Using the atomic number to order the matrix (see left panel of Figure 7.15), i.e. taking into account the similarity of elements along a period of the periodic table, improves this value to $\mathcal{F} = -7.68$. Using the Pettifor scale (right panel of Figure 7.15) yields $\mathcal{F} = -15.47$. As we can see this an excellent improvement. This value decreases further to $\mathcal{F} = -15.62$ by eliminating the elements for which there are no entries in ICSD. Finally, the optimal ordering coming out of our genetic algorithms (see Figure 7.16) yields $\mathcal{F} = -16.91$.

Not only our approach (labeled "GA" in Table 7.1) led to a better quantitative results,

but we can clearly see a significant qualitative improvement of the matrix. Now, in the upper right corner of Figure 7.16 we can clearly identify two blocks (darker squares) that represent subgroups of chemical elements that are much similar between themselves than to any other element of the periodic table (we stress that this structure was absent from Pettifor's original scale). The first of these blocks includes the elements H, F, Cl, Br, I. It is interesting to notice that H appears together with the halogens, which is justified by the fact that most of the H substitutions present in ICSD are with F. This supports the argument that H, F, and Cl form triad, and that should therefore be placed in the same group of the periodic table [155].

Then comes another group containing O, S, Se, Te, Po, Bi, Sb, As, P and N, with a clear subgroup formed by the chalcogens. The next group contains only C and B, which are known to be somewhat special elements of the periodic table. Then there is a group of 9 elements containing Be and the remaining members of the boron group (Al, Ga, In, and Tl) and of the carbon group (Si, Ge, Sn, and Pb). We note that this group is considerably less well-defined than the previous ones, with several possible substitutions outside itself. The next group is constituted by transition metals (plus Mg), followed by the actinides and then by the lanthanides. The end of the table is quite well defined and it is essentially unchanged from the original Pettifor scale. It includes the rest of the alkali earth metals (Ca, Sr, Ba, and Ra), then the alkali metals (Li, Na, K, Rb, and Cs), and the noble gases (Kr and Xe). The radioactive rare-earth Pm appears in between the alkali metals and the noble gases, which is strange chemically, but can be understood due to the very small number of entries in ICSD containing this element, leading to very poor statistics. Finally we find all elements for which there is no entry in ICSD, namely He, Ne, Ar, At, Rn, Fr, Es, Fm, Md, No, and Lr.

Our GA scale follows mostly the order of the groups of the periodic table, but there are several cases where it follows the period, or even a diagonal. Note that a relationship is well-known to exist between certain pairs of diagonally adjacent elements, as trends moving down the periodic table are usually the exact opposite of the trend moving across. From the significant diagonal relationships known to exist (Li/Mg, Be/Al, and B/Si), only Li/Mg is not present in our scale.

We have one last task left in order to have a complete chemical scale similar to the one of Pettifor: To reorganize the elements for which there is little or no data in ICSD. We performed the following operations: (i) we restored the normal ordering of the noble gases (He, Ne, Ar, Kr, Xe, Rn); (ii) we inserted At next to I; (iii) we moved Pm to between Nd and Sm; (iii) as the statistics for the actinides is very limited, we restored their normal (atomic number) ordering; (iv) Nb and Ta turn out to be basically interchangeable without changing significantly the value of the objective function. Therefore, we decided to swap their positions to restore the normal group ordering.

Our final modified Pettifor scale (P_m) is given in given in Table 7.1 and the corresponding matrix is shown in the right panel of Figure 7.16.

7.6.4. Conclusion and Outlook

In conclusion, we performed a statistical study of the possible substitutions of a chemical element A by another B in all known crystal structures. This was possible by using a data mining approach performed on the inorganic crystal structure database. With these data we constructed a function P_{AB} that quantifies the chemical similarity between the elements A and B. We showed that the structure of the periodic table of elements can be reconstructed from a visual inspection of P_{AB}, in a similar manner as in Mendeleev's original work (solely based on statistics, and before the discovery of quantum mechanics).

Having access to a measure of chemical similarity, we were able to propose a mathematical construction for a one-dimensional chemical scale, analogous to the famous Pettifor scale, where similar elements are found in neighboring positions. However, and in contrast with the original Pettifor work, our scale encompasses all available information on the crystal structure of materials, and not only on binary phases.

We believe that our proposed "modified Pettifor scale" can be of use not only for the representation of structure maps, but also as a tool for both theorists and experimentalists to study possible chemical substitutions in the quest for new materials with tailored properties.

7.7. Summary

In this chapter, the library of materials available to our high-throughput method (HTM) has been established. A short description of the original data source was presented in section 7.1, establishing the concept of the unique identifier `coll_code` for materials in our dataset.

Criteria to exclude certain entries from our search were then introduced in section 7.2.

In the next section 7.3, a statistical analysis on the remaining set of *usable* materials, where our HTM computational experiments can be performed on, has been given.

The final section presented the materials of the *dataset* generated during our high-throughput search, i.e. the subset of ICSD where calculations have been performed on.

8. Descriptors of superconductivity within the dataset

The descriptors of superconductivity (chapter 5) have been computed for a **dataset** of 8.071 materials (details are given in section 7.4).

Some of these descriptors can be seen as trivial and represent exclusion criteria for superconductivity (see section 8.1). On the other hand, nontrivial descriptors (like the Fermi bond localization (introduced in section 5.3) need to be evaluated on the material dataset, optimized and tested against known superconductors and non-superconductors. This latter point is the main aim of this chapter.

8.1. Trivial exclusion criteria

8.1.1. Magnetic materials

We exclude magnetically ordered materials from the dataset, as magnetism and superconductivity are competing orders, as discussed in section 5.1.

Computing the lowest energy magnetic state of a material is a challenge per se. However, we use the heuristic assumption that whatever the true magnetic ground state of the system is, if a system would have a magnetic ground state of any configuration within LDA, then a ferromagnetic solution is in any case energetically more stable than the nonmagnetic solution. Therefore we assume that the system will be magnetic and not superconducting if we can find a (meta) stable ferromagnetic solution, as magnetic condensation energies are usually larger than superconducting ones. Leaving collinear, ferromagnetic spin polarization as the only magnetic degree of freedom, while neglecting any effect of spin-orbit coupling, yields computational experiments compatible with our high-throughput search. In order to reduce computational cost, such calculations are only performed if d or f transition metals are present in a system. Materials only consisting of main group elements are treated without a spin degree of freedom, as few such materials have spin-polarized ground states; KO_2, briefly discussed in subsection 5.4.1, was such an exception.

Spin polarization in a material is evaluated via a projection m^I_{at} of Kohn-Sham states, represented in a plane wave basis, onto local atomic orbitals as given by the pseudopotential approximation (5.2). We use $m^{\max}$, the maximal value observed on any atom as an indicator of the overall magnetic properties of a system. Figure 8.1 displays a histogram of $m^{\max}$ over the subset of materials containing transition metals (as the spin polarization is only computed for those). An analysis of separate subclasses confirms the expected result that the largest polarizations involve f and second to this $3d$ transition

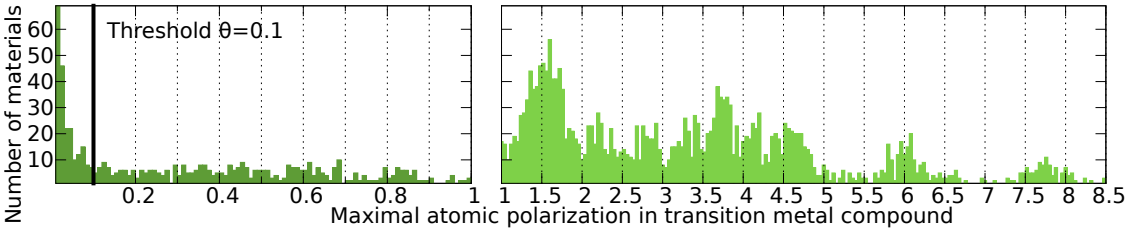


Figure 8.1.: Maximal atomic spin polarization per transition metal compound: representation as a histogram. Width of intervals is 0.01 (left) and 0.04 (right).

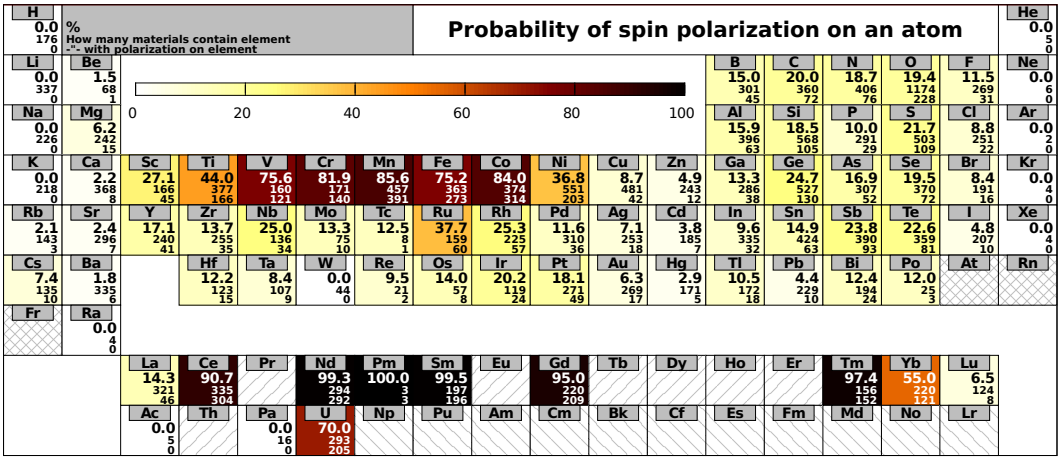


Figure 8.2.: Probability of observing a spin polarization $m_{\mathrm{at}}^{I} > \theta_{\mathrm{at}}$ on an element within a dataset material ($\theta_{\mathrm{at}} = 0.1$, see text).

metals and that essentially all magnetism is correlated with their presence.

The central purpose of analyzing the magnetic propreties within a high-throughput search for superconductors is that we want to *avoid* magnetic materials. Due to possible numerical inaccuracy, comparison with a threshold $m^{\max} > \theta_{\mathrm{at}}$ indicates if a system should be classified as magnetic. Using the detailed histogram in the left panel of Figure 8.1 justifies $\theta_{\mathrm{at}} := 0.1$ as a threshold to separate truly magnetic systems from numerical artefacts, as this region represents the lower flank of the zero-magnetization peak. Employing this threshold, the absolute number of materials, where a given element is classified as spin-polarized can be determined, implying the exclusion of such materials. This information is displayed in Figure 8.2 as a relative frequency or probability for each element. In this representation, it is evident that more than 70% of the lanthanide and actinide and almost 60% of the 3d transition metal compounds in our dataset are considered magnetic; the polarization observed on *p*-block elements is related to their environment.

Altogether, 2.897 transition metal compounds are excluded as potential superconductors by this criterium, which corresponds to 46% of this class, and 36% of the whole dataset.

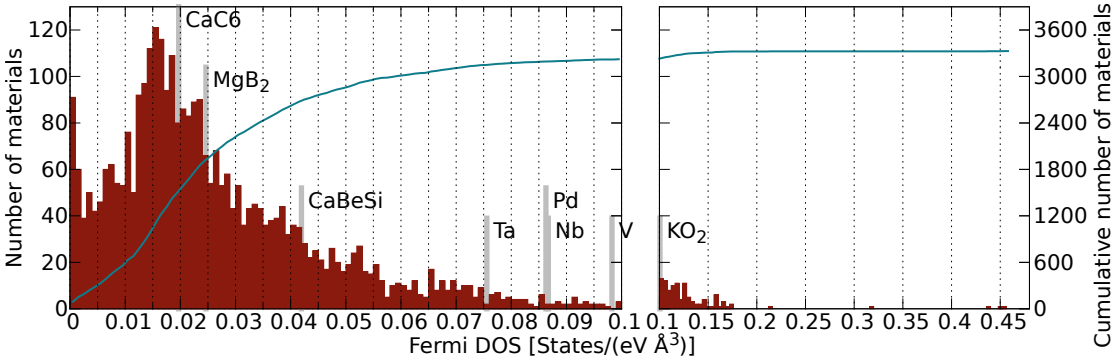


Figure 8.3.: Histogram of DOS_F among nonmagnetic metals. DOS_F for a set of materials has been marked in order to establish the meaning of the scale and a quality of ranking. Width of the histogram bins has been adjusted by a factor of 4 in the right panel.

8.1.2. Insulators

As superconductivity can only occur in *metals*, all materials considered insulators by the criteria defined in section 5.2, cannot possibly exhibit superconductivity. The single-particle spectrum within Kohn-Sham DFT, when the L(S)DA exchange-correlation functional is used, has the tendency [156] to exhibit smaller band gaps than those observed in the experiments, or even exhibit a metallic ground state, while the experimental one is insulating. With these limits in mind, we can apply the aforementioned method to exclude KS-LSDA insulators from further processing.

Appropriate *doping* by holes or electrons can introduce partially filling to the localized, covalent bonds frequently found in insulators. As shown in section 5.3, such localized bonds at Fermi level greatly enhance the strength of the electron-phonon coupling, which in turn enhances the transition temperature T_c. Therefore future application of the method developed within this work will also span doped insulators, or even focus on such systems.

At present, this descriptor implies the expected exclusion of all noble gas compounds from further considerations, as well as 54% of oxygen- and 63% of the halogen compounds. In total, there were 1.846 insulators detected within the dataset, which cannot be superconductors unless appropriately doped, and are therefore excluded.

8.2. Density of states at the Fermi level

The first descriptor used for ranking candidate materials according to their expected superconducting properties is the density of states at the Fermi level (DOS_F). As has been explained in section 5.2, electron-phonon scattering processes are, due to the energy scale of phonon energies, only possible between phonons and electronic states at the Fermi level, and the DOS_F, by definition, describes the availability of such states in a given system.

Throughout this work, in order to ensure the comparability of numerical values be-

Material	CaBeSi	Pd	Ta	V	Nb	CaC_6	MgB_2
T_c [K]	0.4	0.0	4.5	5.4	9.3	11.5	39.0

Table 8.1.: Superconducting critical temperatures of the materials marked Figs.8.4, 8.3

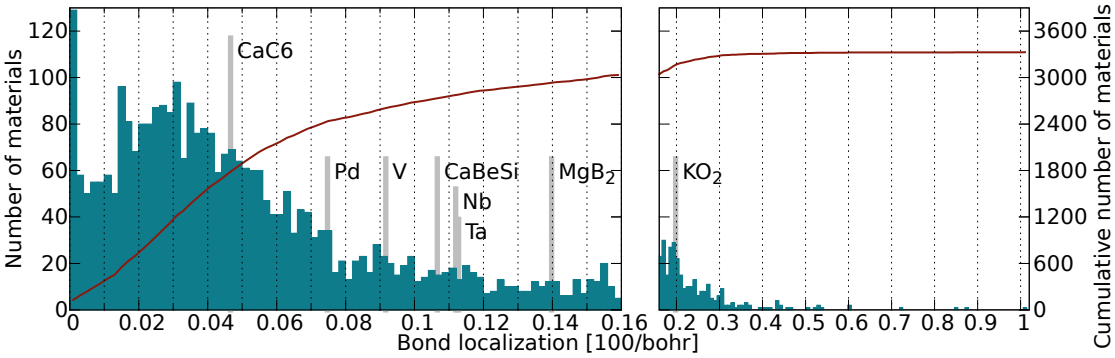


Figure 8.4.: Histogram of bond localization b_F among nonmagnetic metals. Values for a set of materials have been marked, such that both scale and ranking quality can be seen. Width of the histogram bins has been adjusted by a factor of 4 in the right panel.

tween the largely different crystal structures within the dataset, we employ a normalization with respect to unit volume for DOS_F, implying the units $[DOS_F] =$ number of states/eV/Å^3 used throughout this work.

A histogram of the DOS_F over the nonmagnetic metals of the dataset is presented in Figure 8.3, with indicators marking the actual DOS_F values for a test set of materials. While the indicators have been mainly included in order to establish a meaning to the scale, they also demonstrate that the DOS_F alone cannot establish a good ranking among the systems with respect to their superconducting properties: CaBeSi, with $T_c \approx 0.4$K, possesses a far larger DOS_F than MgB_2 ($T_c \approx 39$K) or CaC_6 ($T_c \approx 11.5$K). Palladium, a non-superconductor and Niobium ($T_c \approx 9.25$K) sort very closely, despite the large difference in critical temperature, and far above MgB_2. Moreover KO_2, as a strongly correlated, molecular crystal poorly described within LSDA Kohn-Scham DFT, exhibits within this approximation one of the highest DOS_F values observed in the whole dataset, due to the very narrow bands at Fermi level (subsection 5.4.1), while it is characterized as an insulator by experiment. A representative sample of the materials in the right panel has been analysed, coming to the conclusion that such high DOS_F values are all related to molecular and strongly correlated systems.

As a first result regarding the actual search for new superconductors, a large number of materials does exist in our dataset according to Figure 8.3, having similar DOS_F as MgB2, the conventional superconductor with the highest known critical temperature.

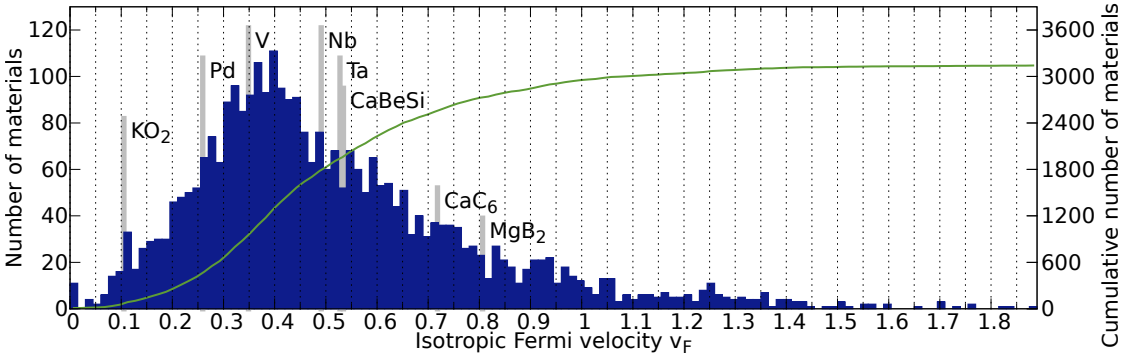


Figure 8.5.: Histogram of isotropic Fermi velocity $\bar{v}_\mathrm{F}$ among nonmagnetic metals. Values for a set of materials have been marked, such that both scale and ranking quality can be seen.

8.3. Fermi bond localization

The Fermi bond localization b_F is our second descriptor of superconductivity employed for a ranking of the nonmagnetic metals according to their expected superconducting properties. Its introduction as a descriptor, as described in section 5.3, is based on the physical properties of MgB_2, and can by the means of a simple model be directly connected to the scattering amplitude of electronic states at the Fermi level, exposed to a unit amplitude bond-stretching deformation of the lattice.

Figure 8.4 presents a histogram of localization values, evaluated via the method described in section A.1, on the nonmagnetic metals of the dataset, which is, like the one presented for DOS_F in the previous subsection, decorated by indicators for b_F values found in a set of example systems.

Also alike the previous case, the highly localized contributions in the right panel can be related, by statistical analysis on the contibuting materials, to systems where strong correlation effects occur.

When investigating the ranking order arising from b_F, palladium and niobium appear well-separated and in the order of their respective critical temperatures (cf. Table 8.1); the same is true for MgB_2 and CaBeSi, which were also the example case originally motivating the introduction of this quantity in section 5.3. CaC_6 does not, however, rank appropriately within *either* DOS_F or b_F derived ordering, a fact that will be discussed in more detail in section 10.2.4, in the context of a collective evaluation of the ranking descriptors.

Regarding our high-throughput search, far fewer materials appear in the b_F-vincinity of MgB_2 than it was in the case of the DOS_F-vincinity, as it lies in the higher tail of the distribution.

8.4. Isotropic Fermi velocity

The last of the ranking descriptors, presented in section 5.4, is the isotropic measure of the Fermi velocity $\bar{v}_\mathrm{F}$, which was originally introduced to detect molecular crystals, a class of materials that has many features in common with strongly correlated systems. Like DOS_F and b_F, also $\bar{v}_\mathrm{F}$ was computed for the whole dataset, but we restrict our analysis to the nonmagnetic metals of the dataset.

In Figure 8.5, the distribution of this dataset among the range of $\bar{v}_\mathrm{F}$ is presented, including markers for the values observed on the example materials. As a first observation, and as mentioned in the discussion in subsection 5.4.1, the Fermi velocity of the moleclular crystal $\mathsf{KO_2}$ belongs to the lowest observed within the dataset. An analysis of the other materials exhibiting similarly low or even lower $\bar{v}_\mathrm{F}$, identified a group of isostructural compounds, with the potassium atom substituted by another alkaline. The chemical composition of many other materials in the $\bar{v}_\mathrm{F}$-vincinity of $\mathsf{KO_2}$ suggests strong correlation, the majority being transition metal oxides. This finding is a strong indicator that $\bar{v}_\mathrm{F}$, evaluated on the Kohn-Sham band structure, is not only a measure for molecular crystals, but also assists in detecting other strongly correlated systems. Furthermore, Pd, where absence of a superconducting transition is related to spin fluctuation, lies within the lower flank of the $\bar{v}_\mathrm{F}$-distribution, between $\mathsf{KO_2}$ and the other example systems.

When evaluating the ranking properties of $\bar{v}_\mathrm{F}$ and more specific, the ordering of the three hexagonal layered structure $\mathsf{MgB_2}$, $\mathsf{CaC_6}$ and CaBeSi, it can be observed that this descriptor indicates the right ordering of the three materials. Actually, it is the only of all three ranking descriptors which would assign a higher rank to $\mathsf{CaC_6}$ than all of the example elementary metals.

Looking at the number of materials in the direct vincinity of $\mathsf{MgB_2}$ in the histogram Figure 8.5, far fewer materials appear in the b_F-vincinity of $\mathsf{MgB_2}$ than it was in the case of the DOS_F-distribution; similar to its position relative to the b_F distribution, $\mathsf{MgB_2}$ lies in the higher tail of all $\bar{v}_\mathrm{F}$ observed in the dataset.

8.5. Statistical correlation among the ranking descriptors

The ranking descriptors DOS_F, b_F and $\bar{v}_\mathrm{F}$ represent different, summarizing views on the electronic structure at the Fermi level. However, on the basis of simple physical arguments, a certain degree of statistical correlation among these quantities is expected: in the extreme case of "flat" bands the corresponding to strongly localized molecule-like electronic states lead also to high DOS_F (cf. section 5.4), which is also demonstrated by the position of $\mathsf{KO_2}$ within the three histograms presented earlier in this section.

One may therefore be concerned about the information content added by the inclusion of a further descriptor, such as the Fermi bond localization introduced as a new concept in section 5.3 or the Fermi velocity section 5.4. This question could be reformulated as: inhowfar can the variation of one quantity be explained by the variation in another? Statistically, given a model for the interdependence of random variables, this question

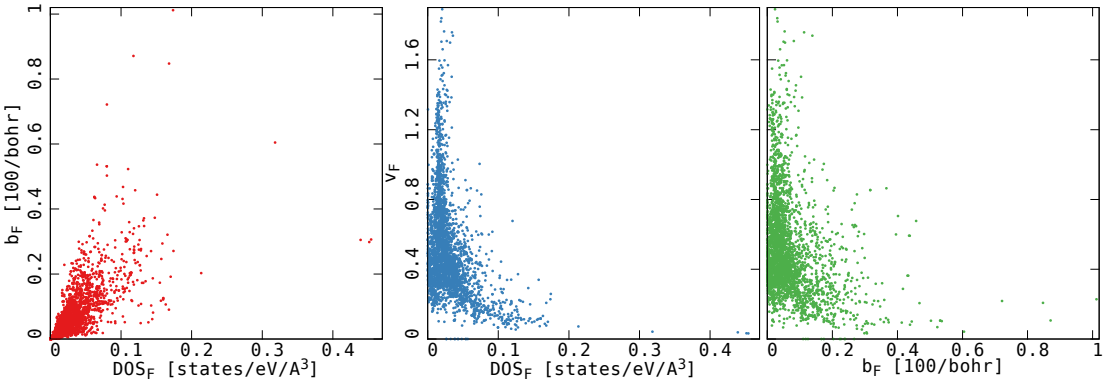


Figure 8.6.: Relation between the ranking descriptors DOS_F, b_F and $\bar{v}_\mathrm{F}$ within the dataset: low statistical correlation between all three quantities is observable (quantitative evaluation is presented in text)

is answered by the *coefficient of determination* r^2 as obtained on a set of data. r is the *Pearson correlation coefficient*, which is defined as

$$r := \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X \sigma_Y}$$

assuming a linear model; $E[\,\bullet\,]$ denotes the expectation, $\mu_{X,Y}$ the means within the dataset and $\sigma_{X,Y}$ the standard deviation of the individual quantities.

An an application of this method to the nonmagnetic metals yields $r^2_{\mathrm{DOS}_\mathrm{F},b_\mathrm{F}} \approx 0.48$, $r^2_{\mathrm{DOS}_\mathrm{F},\bar{v}_\mathrm{F}} \approx 0.13$ and $r^2_{b_\mathrm{F},\bar{v}_\mathrm{F}} \approx 0.08$. Less formally speaking, this means that assuming a linear model, variation of DOS_F could explain only about 50% of the variation of b_F, and both b_F and DOS_F could explain only 10% of the variation of the Fermi velocity mean $\bar{v}_\mathrm{F}$. In Figure 8.6, scatterplots of descriptor pairs are presented; aside from the linear model rejected by the coefficient-of-determination analysis, the three subfigures emphasize that no other model could explain an interdependence of the descriptors. Only fairly general trends can be observed: DOS_F could, at most, provide information about an envelope function to b_F, i.e. an upper and lower bound. Furthermore, the hypothesis of a correlation between low $\bar{v}_\mathrm{F}$ and high DOS_F or b_F is confirmed by the middle and right panel of Figure 8.6: very high values of the latter two quantities are actually only observed within the range of low $\bar{v}_\mathrm{F}$ (note that this does *not* mean that low-$\bar{v}_\mathrm{F}$ would *imply* high DOS_F or b_F).

The result is actually encouraging, as it demonstrates that all three ranking descriptors significant information inexplicable by the others. Without going further into *information theory*, this simple finding illustrates that the information provided by triples $(\mathrm{DOS}_\mathrm{F}, b_\mathrm{F}, \bar{v}_\mathrm{F})$ of all three descriptors exceeds that of any pair of descriptors (or the individual descriptors by themselves).

This is an important finding for the introduction of a collective evaluation scheme, presented in chapter 10.

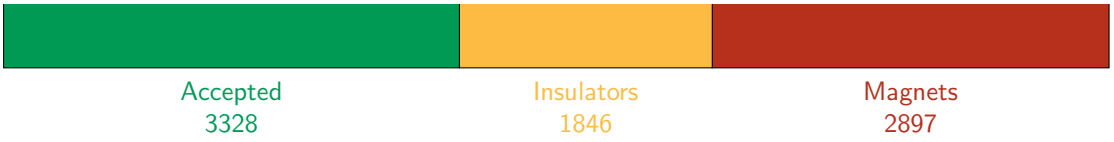


Figure 8.7.: Exclusion of insulators and magnets

8.6. Summary

In this chapter, we have presented an evaluation of the *descriptors of superconductivity* within the dataset obtained during our high-throughput search for superconductors. As a first step, we excluded materials with unwanted properties, such as magnets or insulators (Figure 8.7), leaving 3.328 (out of 8.071) nonmagnetic metals.

The second set of descriptors, the *ranking descriptors*, were applied to the nonmagnetic metals. The ranking properties regarding a small set of superconductors have been described: while each of the descriptors did order a different subset of the structures in agreement with the superconducting critical temperature T_c, none of the descriptors by itself could establish the correct order for the whole example set. A brief analysis of the statistical correlation between the ranking descriptors showed them to be largely independent from each other, meaning that each of the descriptors does provide information inexplicable by the other two.

In the next chapter, we continue our investigation of high-throughput methods by evaluating the feasibility to *predict* the descriptors of superconductivity directly from the crystal structure of a material, via the machine learning techniques described in chapter 6.

An evaluation *combining* all ranking descriptors is presented in chapter 10.

9. Prediction of electronic properties via machine learning

Machine learning (ML) methods, introduced in chapter 3, have the potential to greatly accelerate a high-throughput search for superconductors, when two essential conditions are fulfilled: an *input vector* representation of crystal structures needs to be found and a sufficient amount of *training data* has to be available.

In chapter 6, we proposed such an input representation of crystal structures. The data obtained in the first stage of our high-throughput search, which was statistically described in the previous chapter 8, can and does serve as a training and test set for our ML experiments.

In this chapter, we present the results of these experiments, evaluating the actual performance of ML methods in combination with our proposed input representation.

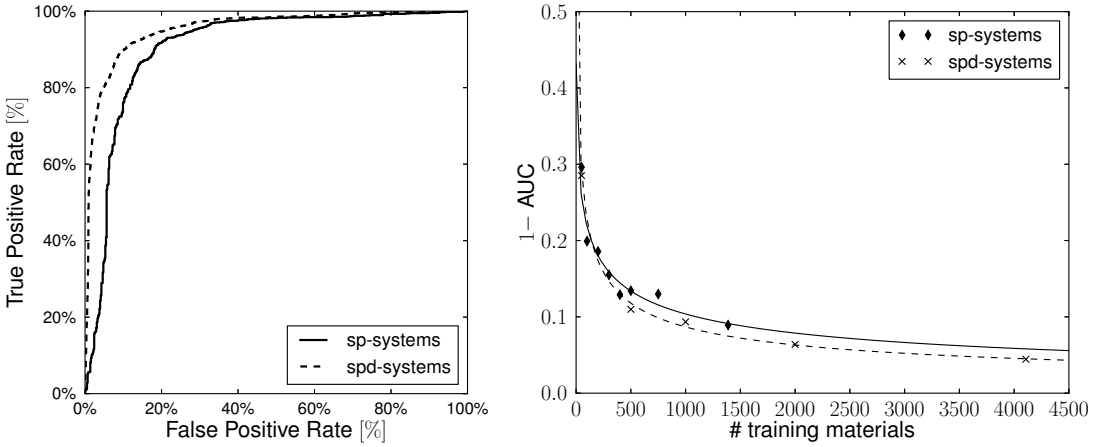
As a first step, a brief description of the ML dataset, a subset of the one presented in the last chapter, is given in section 9.1.

The following two sections evaluate the predictions made by a support vector machine (SVM) metal/insulator classifier (section 9.2) and kernel ridge regression (KRR) predictor for $\mathrm{DOS_F}$ (section 9.3).

9.1. Machine Learning Dataset

In our ML experiments, a subset of the materials presented in section 7.4 is employed. The chosen subset contains only non-duplicated materials with a maximum of 6 atoms per primitive cell. We subdivide the set into *sp* (1716 crystals) and *spd* (5548 crystals), while we explicitely exclude any material containing f transition metals. The latter choice is based on the limited reliability of our Kohn-Sham calculations with the chosen computational parameters in such systems, which from the perspective of ML algorithms manifests itself in essential *noise*, i.e. almost contradictory information contained in the *labels* of the training and test set.

We use nested cross-validation (subsection 3.2.2) for the model selection process [157, 158], i.e., the parameter selection and performance evaluation are performed on separate held-out subsets of the data that are independent from the set of training materials. This ensures to find optimal parameters for the kernel and the model regularization in terms of generalization while avoiding overfitting.



(a) Receiver operating characteristic (ROC) curve: dependence of sensitivity and specifity on the choice of the classification threshold

(b) Learning curve: the cumulative error is measured by the area under the ROC curve (AUC)

Figure 9.1.: Metal-insulator classification by an SVM, employing a Gaussian kernel, with input in PRDF representation

9.2. Metal-Insulator classification

The first and foremost important descriptor of superconductivity is the metallicity of a system, and, due to the magnitude of the band gap in relation to typical phonon energies, no superconductivity whatsoever could occur in semiconductors or insulators. In our predictions, no explicit distinction is made between the latter two cases, as we label every material with a finite band gap in the Kohn-Sham spectrum as an insulator.

Distinguishing metals from insulators poses a *classification problem*, to which we apply a *support vector machine (SVM)* that finds a separating hyperplane in feature space, while maximizing the space between the two classes [86]. Crystal structures are represented as PRDF feature matrices (Equation 6.1), in conjunction with a Gaussian radial basis function kernel (Table 3.1). Training and tests have been performed on the full *spd* dataset, and in addition on the *sp* subset. A perfect separation of insulators from metals be acheived in neither case. By shifting the classification threshold, sensitivity vs. specifity, i.e., the trade-off between correctly detected insulators and metals incorrectly classified as insulators can be adjusted. For example, our classifier is able to detect 85.0% of the insulators while only mistaking 7.3% of the metals as insulators on the whole *spd* data set. This tradeoff is characterized by the limiting curve, termed *receiver operating characteristic* (ROC), and is represented in Figure 9.1a. The area under the ROC curve (AUC) provides a measure for the cumulative error, as it approaches unity in the case of a perfect classification. In Figure 9.1b, the *learning curve* of our SVM predictor is presented, which indicates how strongly the cumulative prediction error, measured as $1 - \mathrm{AUC}$, reduces with the size of the training set. A monotonical decrease of error with

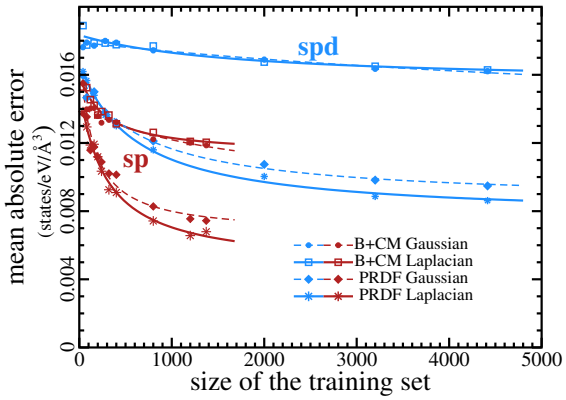


Figure 9.2.: Learning process as a function of the number of training materials for both feature representations, kernels and datasets.

Predictor Kernel	Features	Systems sp	Systems spd
Mean	–	1.50±0.02	1.82±0.03
KRR			
(linear)	B+CM	1.45±0.04	1.68±0.01
(gauss.)	B+CM	1.19±0.03	1.62±0.01
(lapl.)	B+CM	1.20±0.04	1.63±0.02
KRR			
(linear)	PRDF	0.87±0.02	1.68±0.03
(gauss.)	PRDF	0.74±0.03	0.95±0.02
(lapl.)	PRDF	**0.68**±0.03	**0.86**±0.01

Table 9.1.: Mean absolute errors and standard errors of DOS predictions in units of 10^{-2} number of states/eV/Å^3

the size of *sp* and *spd* training sets can be observed, with an $\frac{1}{x}$-like behaviour very similar in both cases; however, as the former represents a small subset of the latter, the precision of the *sp* predictor is significantly worse than in the *spd* case, as can be observed in the limiting ROC curves in Figure 9.1a. The asymptotic behaviour, however, suggests that any significant increase of prediction accuracy would imply the necessity of increasing the size of the training set by a factor of at least 2, which is left as a future exercise.

9.3. Fermi density of states as a regression problem

The prediction of continuous labels, such as the numerical value of the density of states at Fermi level (DOS_F), poses a *regression problem* in ML terminology. We employ *kernel ridge regression* (KRR, section 3.4.2) for this prediction, which also provides an estimate of the *predictive variance* which can serve as a measure of how well a material of interest is represented in the training set.

As in the case of the metal/insulator classification, the full *spd* material set and its *sp* subset are considered separately for the DOS_F regression.

The mean absolute errors of the predictions of the two crystal representations outlined in chapter 6 are collected in Figure 9.1, evaluated considering three different kernel functions: the linear kernel corresponds to a direct evaluation of the inner products in input space, while Gaussian and Laplacian kernels perform an implicit mapping to the corresponding radial basis function (RBF) feature spaces Table 3.1 Furthermore, we list the mean predictor, which always predicts the average DOS_F value of the training set, as a simple baseline. Both representations yield models that are significantly better than the mean predictor.

Figure 9.2 illustrates how the error decreases steadily with increasing number of materials used for training, with the mean absolute error (MAE) as an error metric, the same which was employed in our KRR optimization. However, the PRDF features con-

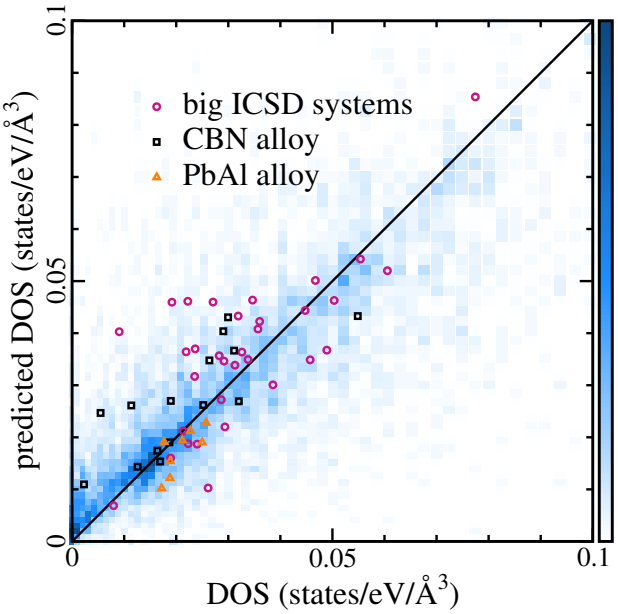


Figure 9.3.: Comparison between predicted and calculated $\mathrm{DOS_F}$ for spd systems. The background density distribution refers to the cross validation systems. Dots are additional systems (see legend) of far larger size than those used for training.

sistently outperform the B+CM description, which is why the further analysis will focus on the PRDF representation, with the slightly better performing Laplacian kernel.

The higher complexity of the spd systems can be clearly observed in the learning curves, which show how much better the prediction problem can be solved as a function of the available data. The mean error is much lower in sp materials. Furthermore, the learning curves are steeper, i.e., increasing the training set size within the restricted materials class improves the prediction accuracy rapidly. One origin of this higher complexity lies in the growing dimensionality of the input space: given N_{el} possible chemical elements in all material compositions, $\dim(\mathrm{X}) \propto N_{\mathrm{el}}^2$. Furthermore, by including materials with d electrons, the physics becomes more rich. Due to both reasons, much more training data is required to achieve an improvement comparable to that of sp systems.

Furthermore, when comparing the learning curve in Figure 9.2 with the metal/insulator classification one Figure 9.1b, the increased complexity of the regression problem, when compared to the classification one, can be clearly observed from the slopes and asymptotic behaviours of the curves.

The cross-validated prediction of $\mathrm{DOS_F}$ for spd systems is shown in Figure 9.3, as a density plot of computed versus predicted values. It is evident that the density is accumulated along the diagonal of the plot, demonstrating that the machine is giving meaningful predictions, while the average error is smaller than 6% of the DOS value range. Thus this result represents another proof of principle that a complex output of the Kohn-Sham equations can be predicted directly by means of machine learning – albeit the considerable variance of errors.

From Figure 9.2 it is clear that, in order to increase the prediction accuracy, the size of the training set should be increased, possibly at the limits of present computing facilities. Instead of a brute-force approach, the problem may become less costly by using an active learning scheme, e.g., by extending the set where the *predictive variance* is higher. We still expect that in order to obtain highly accurate results the computational cost will be large. Can the proposed approach still be useful at the present accuracy level?

To answer this question we first point out that ML is at least *2 or 3 orders of magnitude faster* than any direct computational approach, this means that a fast ML scan may always be of use as a preliminary step before consuming precious resources on detailed calculations. Second, a remarkable feature of the PRDF representation is that it is not fixed to a certain number of atoms in the unit cell during training and prediction. This means that once the predictor has been trained, it can be used to predict the properties of any other system. This is virtually independent from its size as long as it is well represented by the training set.

As a proof for this ability, we consider additional test systems, divided into 3 set: The first set (pink circles in Figure 9.3) contains only systems taken from the ICSD, chosen among those with between 30 and 80 atoms per unit cell that are well represented by the training set. Therefore, only ICSD materials with a relatively low predictive variance were chosen for calculation. The second set (orange triangles in Figure 9.3) contains a purely metallic alloy of lead and aluminum. All the systems in this set are crystals with 125 atoms per unit cell, differing by the Al/Pb concentration. The third set (black squares in Figure 9.3) is a solid solution of three atomic types in a diamond lattice: Carbon, Boron and Nitrogen, at different composition and a total of 45 atoms per unit cell.

Unlike the training systems, each of these involve a large computational cost, and would not be feasible without access to a computation facility. While the PbAl alloys are quite well predicted, some of the $\mathrm{DOS_F}$ for the large ICSD systems (mostly oxides) are overestimated, as well as some of the $\mathrm{DOS_F}$ of the CBN solid solution. Again, a clear diagonal accumulation is achieved. Nevertheless for these large systems, which the predictor has never been trained on, the average quality of the prediction of large systems is comparable to that of the much smaller, cross-validated systems.

9.4. Summary

In summary, we have investigated a machine learning approach for fast solid-state predictions. A set of LSDA calculations has been used to train predictors of both metallicity and $\mathrm{DOS_F}$. We expect that our method can be extended to directly predict other complex materials properties as well. It certainly can be combined with other, more accurate, electronic structure techniques such as *GW*.

The accuracy of predictions depends strongly on how crystals are represented. We found that Coulomb matrices, while being successful for predicting properties in small organic molecules [127, 128] belonging to a, by the choice of constituent elements, small subspace of chemical compound space, are not suitable to describe crystal structures in

the comparatively large, coarsely sampled subspace relevant to this work.

Instead, we have proposed a representation inspired by partial radial distribution functions which is invariant with respect to translation, rotation and the choice of the unit cell.

Our results clearly demonstrate that a fast prediction of electronic properties in solids with ML algorithms is indeed possible. Although presently the accuracy leaves room for improvement, we consider the predictions useful for a first screening of a huge number of materials for properties within a desired value range. In a second step, high-accuracy electronic structure calculations are then performed on the promising candidates only. What makes the approach extremely appealing is that the PRDF representation allows to learn on small systems with low computational cost and then extrapolate to crystals with an arbitrary number of atoms per unit cell, for which conventional DFT calculations would be prohibitive.

9.5. Outlook

The presented machine learning approach to the prediction of electronic properties of crystal structures is definitely promising, while the presently achieved accuracy, evaluated by metal/insulator classification and $\mathrm{DOS_F}$ regression, still lies below the reliability required in a high-throughput search for superconductors; improvement is left open for future work.

For one, increased accuracy could be achieved by a significantly enlarged training set, which could span also non-equilibrium structures.

The choice of representing periodic elements as *dimensions* within PRDF representation opens a path to develop an improved metric considering true *chemical* similarity of the periodic elements.

Moreover, the PRDF representation can be extended to encode also local, *angular* information by switching from an element-*pair* to an element-*tuple* (in the next-lowest order, *triplet*) based approach. Given that such an approach was successful, and given that non-equilibrium configurations were present in the training set, the applicability of our method could be greatly extended, including the use-case of (conventionally computational expensive) structural relaxion, where forces and stress could be provided by a ML predictor.

10. Prediction of superconductivity from the descriptors

The descriptors of superconductivity (chapter 5) have been introduced to provide a computationally cheap estimate of the superconducting properties of a system. They have been computed for a **dataset** of 8.071 materials (section 7.4), with the obvious intention to identify or predict superconductors. In chapter 8, both magnets and insulators have been removed from the dataset, leaving 3.328 nonmagnetic metals as possible superconductors. The descriptors were then tested for their individual predictive capabilities.

In this chapter, we introduce a prediction scheme for superconductors, based on all descriptors, and establish its predictive power both on known superconductors and non-superconductors, and on a sample of the dataset.

A statistical analysis then examines relations between crystal structure, chemical composition and predicted superconductivity, inspired by *Matthias' rules to find new superconductors*.

Some basic information on the predictied superconductors is collected in the final section 10.4.

10.1. Descriptor Validation Set

The descriptor validation set is a set of materials with *known* superconducting properties and descriptors of superconductivity. It can therefore be used to evaluate the quality of collective evaluation methods. Moreover, the relation between the descriptors of superconductivity and the calculated superconducting properties of these materials serves as an empirical basis to the classification scheme introduced later within this chapter.

In order to be suitable to this goal, the validation set needs to contain both

non-superconductors, which in the context of this work are defined as systems where a phonon-mediated pairing mechanism could not lead to $T_c > 1$K, or alternatively exhibit an electron-phonon interaction parameter $\lambda < 0.34$

and (phononic) superconductors, having as many different chemical compositions and crystal structures as possible.

Moreover, just out of curiosity, a small set of materials is included where the experimentally observed T_c could *not* be explained by phononic mechanisms, such as layered ferropnictides and cuprate high-temperature superconductors. While the mechanism is non-phononic, as long as the electronic states are well described within the Kohn-Sham picture, and the energy scale of the interaction does not strongly differ from the phononic

Material	λ	T_c [K]
MgB_2	0.64	18.4
CaC_6	0.90	20.1
V_3Si (A15)	1.43	28.3
ZrN	0.75	15.8
TaC	0.66	9.2
Al	0.43	2.0
Sn	0.84	5.1
Nb	1.34	16.1
Pd	0.34	0.19

(a) Widely known materials

Material	λ	T_c [K]
$RbGe_2$	1.64	10.3
$BaSi_2$	1.56	20.8
CaBeSi [40]	0.37	0.9
BeB_2	0.24	0.0
FeSe	0.07	0.0
LiFeAs	0.18	0.0
$YBa_2Cu_3O_7$ (YBCO7)	0.00	0.0

(b) Materials supplied by other projects within our research group

Table 10.1.: Example materials from the descriptor validation set. Non-phononic superconductors are highlighted in blue and materials with $T_c < 1\,\mathrm{K}$ in red.

one, Fermi bond localization b_F and Fermi density of states DOS_F may be considered in a more general way: b_F is a measure for the electronic scattering amplitude under *any* mechanism, which could be expressed as a perturbation of the self-consistent potential, localized within the bonding region; DOS_F, on the other hand, is a measure for the electronic phase space available for scattering processes.

10.1.1. Computational considerations

Within this work, superconductivity is treated at the level of isotropic Eliashberg theory in the Allen-Dynes approximation (subsection 2.2.6, (2.39)). The critical temperature is determined by the isotropic electron-phonon interaction parameter λ and isotropic weighted averages $< \omega^2 >$, $\omega_{\log}$ of the phonon frequencies. Coulomb repulsion is accounted for by the Morel-Anderson coulomb pseudopotential μ^*, which is conventionally fitted to reproduce the experimentally observed T_c. Despite the simplification implied by an isotropic treatment, the computational complexity is still high, and dominated by the linear-response phonon calculations (section 1.2) necessary to obtain λ, $< \omega^2 >$ and $\omega_{\log}$. Evaluating T_c for a single material demands *weeks* of researchers time. In consequence the validation set is comparatively small, consisting of 75 materials. Furthermore, it is statistically biased, as some classes of materials are overrepresented. Moreover, as the number of phonon modes and therefore the computational complexity of electron-phonon calculations scales as $\mathcal{O}(N_{at}^3)$ with the number of atoms per primitive cell N_{at}, materials with *small* unit cells dominate the validation set. In consequence, simple structures are strongly represented, potentially limiting the variety of superconducting behaviour observable within the validation set.

Material	coll_code	λ	T_c [K]	Material	coll_code	λ	T_c [K]
$AuTe_2^{HP}$ [160]	66626	1.22	4.2	KO_2	38138	0.53	0.7
BC_5 [161]	166555	0.83	43.0	Ag_2O	20368	0.50	2.0
Be_2B [162]	20384	0.66	19.8	CaIn	619377	0.52	1.2
$InAs^{HP}$ [163]	43972	0.83	6.8	$CuAlS_2^{HP}$	165739	1.23	25.7
PbS^{HP} [164]	77865	1.65	22.3	$CuAlSe_2^{HP}$	165741	0.90	9.9
$PbSe^{HP}$ [165]	77870	1.59	12.4	$CuAlTe_2^{HP}$	165743	0.43	1.4
PIn^{HP} [166]	53104	1.13	16.1	LiPb	104762	0.78	3.2

Table 10.2.: Descriptor validation set: example ICSD non-magnetic metals, spanning non-superconductors and superconductors, as predicted by our descriptors of superconductivity (see text). Non-superconductors are highlighted in red. In case a scientific reference could be found, the respective table row is printed in black and includes the citation; remaining, prospective superconductors are highlighted in green. In all cases, the ICSD collection code (unique identifier within ICSD) is reported, describing the material *prior* to the structural relaxation performed within the present work.

10.1.2. Origin

Due to the high computational cost, an initial set of materials was assembled by Antonio Sanna from materials studied in previous projects and unpublished data.

Data from previous projects cover widely known phononic supterconductors, such as the hexagonal layered structures MgB_2 and CaC_6, zirconium nitride ZrN and members of the A15 phases, such as as V_3Si or Nb_3Sn. Moreover, a set of elemental superconductors, such as Al, Pb or Nb, and some high-pressure phases of metalloids and nonmetals has been included. A subset of this class is presented in Table 10.1a.

Less widely known materials provided by current research projects cover different high-pressure phases of S, the families of layered digermanides and disilicides [159], and layered materials, such as CaBeSi and BeB_2, which served as additional reference points in the analysis of the superconductivity in MgB_2.

The superconductivity prediction scheme, which will be described in section 10.2, had been parametrized with the initial materials. It was then applied to the nonmagnetic metals found by our high-throughput calculations (cf. section 8.1). Materials were selected on the basis of these predictions, and explicit calculations of λ and T_c were made. As the number of non-superconductors within the initial set was low, we intentionally included a set of predicted non-superconductors, testing the scheme for false negative predictions, meaning superconductors predicted as non-superconductors. The predictor was refined based on the new data, iteratively improving the quality of predictions based on the findings within groups of new materials added to the validation set.

In order to further explore our hypothesis of DOS_F, b_F and $\bar{v}_F$ being descriptors of superconductivity, materials with large (in relation to the initial materials) individual values in each of the three have been included: following our hypothesis, unless significant values are present in the respective other two descriptors, such materials are expected to be non-superconductors (example materials are marked in red in Table 10.2).

Furthermore, materials have been included that lie within *badly sampled regions*, with $(\mathrm{DOS}_\mathrm{F}, b_\mathrm{F}, \bar{v}_\mathrm{F})$ far from previous materials.

Materials that were found to be dynamically unstable were discarded: the present work is concerned with the prediction of superconductivity in existing structures, and other methods for the prediction of crystal structures do exist [134–138].

As the database used during our research (ICSD) was built from scientifically published data, no implicit claims on the novelty of superconductivity in such materials can be made; scientific references regarding superconductivity for predicted superconductors were frequently found. In Table 10.2, a subset of the materials introduced by our search is presented, marked by black (if scientific references to superconductivity in the material could be found) and green (in case that no conclusive reference could be found).

10.2. Prediction scheme

After having defined the descriptor validation set, we will now introduce a simple prediction scheme for superconductivity, based on the descriptors density of states at Fermi level DOS_F (section 5.2), Fermi bond localization b_F (section 5.3) and isotropic Fermi velocity $\bar{v}_\mathrm{F}$ (section 5.4). We base our considerations on the isotropic, mass-independent electron-phonon interaction parameter λ (2.12).

10.2.1. Importance of Fermi velocity

Special attention must be paid to the Fermi velocity $\bar{v}_\mathrm{F}$, as it was introduced as a measure to detect systems which are expected to be poorly described within LDA Kohn-Sham DFT due to stronger electronic correlation effects.

However, a strict separation of well- and poorly described systems by the means of $\bar{v}_\mathrm{F}$ cannot be defined; therefore, $\bar{v}_\mathrm{F}$ could not be used as a trivial exclusion criterium (section 8.1), while, on the other hand, no theoretical justification for its impact on λ was presented. On this basis, we decided to include $\bar{v}_\mathrm{F}$ simply as an additional *dimension* in the collective evaluation.

10.2.2. Fermi DOS and bond Localization

In chapter 5, both density of states at Fermi level (DOS_F) and Fermi bond localization (b_F) were introduced as descriptors of superconductivity, joined by a discussion of their relation to the isotropic electron-phonon coupling parameter λ. We use the well-known relation $\lambda \propto \mathrm{DOS}_\mathrm{F}$ (section 2.3.2). In section 5.3.1, a connection between b_F and the matrix elements of the electron-phonon interaction was made, concluding a monotonously rising dependence of the deformation potential, i.e. the scattering amplitude under unit deformation and b_F, which was subsequently quantified (Figure 5.3, right panel) within a simple model system. However, the corresponding scaling function within Kohn-Sham systems is unknown, and a good approximation would require detailed data for a training set far larger than the validation set. Using a linear approximation, our estimate is

simply the product

$$\mathrm{DOS}_\mathrm{F} \cdot b_\mathrm{F},$$

which was sufficient for our predictions on the validation set.

10.2.3. Graphical representation

These considerations suggest also a scheme for the graphical representations used during our analysis of $\mathrm{DOS}_\mathrm{F} \cdot b_\mathrm{F}$: essentially, $\mathrm{DOS}_\mathrm{F} \cdot b_\mathrm{F}$ and $\bar{v}_\mathrm{F}$ require independent axis, the same is true for λ, which is supplied by detailed calculations performed outside the framework of out descriptors, and has, due to this, the character of a *label*. Therefore, materials will be presented in a two-dimensional scatter plot, with each point representing a material $\boldsymbol{m}_i$. The coordinates of each point are determined by $\mathrm{DOS}_\mathrm{F}(\boldsymbol{m}_i) \cdot b_\mathrm{F}(\boldsymbol{m}_i)$ and $\bar{v}_\mathrm{F}(\boldsymbol{m}_i)$, while information about λ_i is provided by the color of the associated point if applicable. As $\mathrm{DOS}_\mathrm{F} \cdot b_\mathrm{F}$ differs by orders of magnitude between different materials, logarithmic axis scaling is applied.

10.2.4. Application to the descriptor validation set

The descriptor validation set is graphically represented in Figure 10.1. While being, as previously discussed, far from a complete description of the electron-phonon coupling or critical temperature T_c, the relation of the simple product of DOS_F and b_F to superconducting properties can be observed.

First of all, most of the best phononic superconductors, such as $\mathsf{MgB_2}$, TaC or ZrN, accumulate towards the top right of Figure 10.1, in the region of both high $\bar{v}_\mathrm{F}$ and $b_\mathrm{F} \cdot \mathrm{DOS}_\mathrm{F}$.

Also known high-temperature, non-phononic, superconductors, including both pnictides and cuprates do lie within the region of high $b_\mathrm{F} \cdot \mathrm{DOS}_\mathrm{F}$, but in regions of low $\bar{v}_\mathrm{F}$, where the Kohn-Sham LDA is unreliable. We will dismiss this fact in the following, as no *predictive* theory for such superconductors does exist, and thus no connection between superconductivity and descriptors based on Kohn-Sham electronic properties can be established.

Soft-phonon driven layered germanides and silicides

A certain class of superconductors is, however, not so well described within this scheme: superconducting layered *germanides*, the most prominent being $\mathsf{RbGe_2}$, with an expected T_c of 10.3 K [159], and layered *silicides*; although λ is large in some of these structures, they appear shifted towards the lower left in the Figure 10.1. However, there is a clear physical interpretation for this shift: the paradigm of superconductivity being driven by the large deformation potential in covalent bonds does *not* apply to these systems.

Instead, superconducting pairing is driven by soft (low frequency) phonon modes. In turn, the magnitude of the electron-phonon matrix elements $g^{n,n'}_{\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q},\nu}$, Equation 1.31, is strongly enhanced by the prefactor $\omega^{-1/2}$, which covers the intrinsic amplitude of

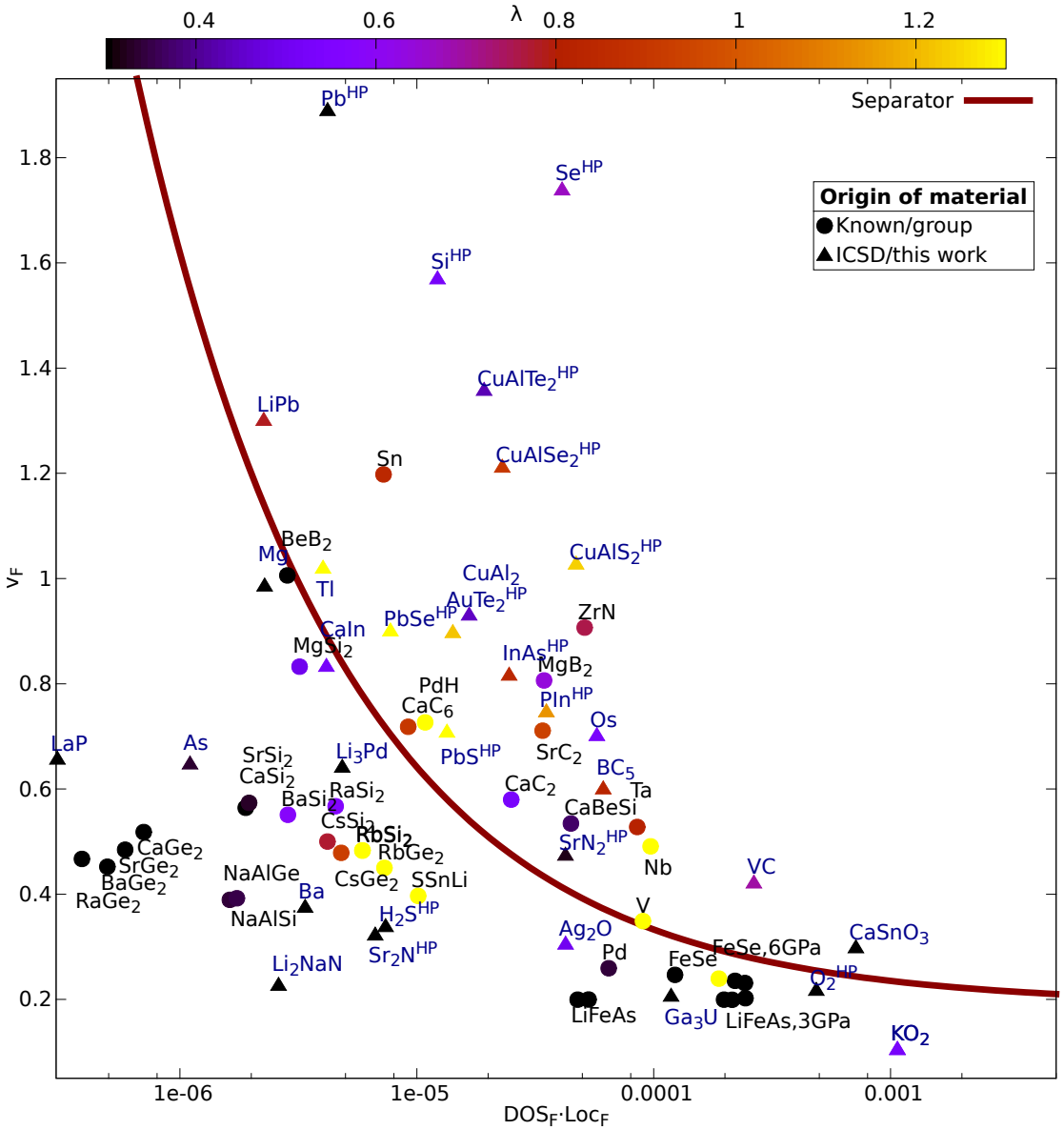


Figure 10.1.: Prediction graph for the Descriptor validation set. Color indicates the value of λ. Good phononic superconductors accummulate in the region of high $\mathrm{DOS_F} \cdot b_\mathrm{F}$, and above the range of $\bar{v}_\mathrm{F}$ indicating strong correlation. The empirically determined separator (10.1) between superconductors and non-superconductors is denoted as a dark-red line.

nuclear motion, while the deformation potential, described by b_{F}, is defined for unit displacement.

Essentially, such soft phonon modes open a second way to achieve superconductivity. But despite the large value of λ, superconductivity is typically weak in such systems, due to the low energy provided by phonons within scattering processes, which seriously restricts the phase space available to phonon-induced electronic transitions (chapter 5). Therefore, such systems are not of *primary* interest when searching for superconductors with a high critical temperature.

However, the problem of our present descriptors observed on this class of systems emphasizes the importance of phonon frequencies. An *improved* set of descriptors should therefore also provide an estimate for such effects, either by predicting the presence of such modes, or by a prediction of the actual magnitude of the frequency-dependent intrinsic-amplitude factor contributing to $g^{n,n'}_{\boldsymbol{k},\boldsymbol{k}+\boldsymbol{q},\nu}$. Given that such a description could be achieved, also systems where soft phonons do provide a *secondary* pairing channel (section 2.2), in presence of a primary, deformation potential driven one.

10.2.5. Predictor

Based on empirical observation within validation set, a *classification curve*, which establishes an approximate border between good phononic superconductors and the remaining systems, has been determined. While, in principle, the problem could also be solved by a machine learning classifier, such as the support vector machine (subsection 3.4.1). within an appropriate feature space, we refrain from such an approach both due to the very limited size of the validation set and the statistical bias.

Observing the distribution of superconductors and non-superconductors within the validation set, we define the simple analytical function

$$s(\mathrm{DOS}_{\mathrm{F}}, b_{\mathrm{F}}) := \frac{a}{\sqrt{\mathrm{DOS}_{\mathrm{F}} \cdot b_{\mathrm{F}}}} + v_{\mathrm{F,min}} \tag{10.1}$$

with the two adjustable parameters a and $v_{\mathrm{F,min}}$ to *separate* superconductors from non-superconductors. The offset $v_{\mathrm{F,min}}$ has been included in order to compensate for the asymptotic $\mathrm{DOS}_{\mathrm{F}} \cdot b_{\mathrm{F}}$ limit, as $\bar{v}_{\mathrm{F}}$ was included as a descriptor in order to avoid molecular and possibly other strongly correlated systems, which may not be well-described within LDA Kohn-Sham DFT. Predictions about if a material will be superconducting are made on the basis of a material's position relative to the function by

$$\Theta(\bar{v}_{\mathrm{F}} - s(\mathrm{DOS}_{\mathrm{F}}, b_{\mathrm{F}})). \tag{10.2}$$

As mentioned in section 10.2.4, a whole class of materials, the soft-phonon driven superconductors, is not well-described, due to the missing estimate of phonon frequencies within our descriptors of superconductivity. During the determination of the model parameters, we do therefore associate far lower weight to all systems within this class.

The results of applying the classification Equation 10.2 to the validation set are presented in Table 10.3. Like in all of this chapter, the labeling of a material as a superconductor or non-superconductor has been solely performed on the basis of the computed lambda.

	superconductors (positive)	non-superconductors (negative)
false	3	12 (8 silicides/germanides)
true	28	30

Table 10.3.: Classification performance of Equation 10.2 within the validation set. The comparatively high number of false negatives (superconductors classified as non-superconductors) has its origin in the low weight of soft-phonon driven systems used during the determination of the model parameters, as we know that this class is both not well-described by our approach and of low practical relevance (section 10.2.4).

The overall *correct classification rate* within the validation set is 79%; the dominant part (50%) of all false classifications has its origin in the false-negatives within the subset of soft-phonon driven germanides and silicides. Without this subset, the correct classification rate is approximately 90%.

The high correct classification rate measured within the validation set is a confirmation of our method, and encourages its application within a high-throughput search. Indeed, a subset of the validation set materials (marked by triangles in Figure 10.1) was actually introduced on the basis of a preliminary predictor (10.1) determined on the basis of the initial validation set materials (marked by circles). The predictor presented here corresponds to that original one, after the application of minor refinement in a and $v_{\mathrm{F,min}}$ based on the datapoints added within this process.

10.3. Application to the high-throughput dataset

We applied the simple predictor (10.2), parametrized on the descriptor validation set, to the nonmagnetic metals of the dataset (Figure 10.2). The predictor excludes 76% of the materials from being good superconductors, as they lie in regions of either too low $\mathrm{DOS_F} \cdot b_\mathrm{F}$ to be superconductors with notable T_c, or too low $\bar{v}_\mathrm{F}$ to be reliably described within LDA Kohn-Sham DFT. Superconductivity is predicted for 750 structures.

This fact is a *major result* for high-throughput methods in the context of superconductivity, as it allows to focus in-depth calculations to a relatively small subset of systems.

From a second, statistical point of view, Figure 10.2 does also provide an indication of why no materials with properties similar to $\mathsf{MgB_2}$ with its exceptional phonon-mediated T_c have been discovered (neglecting the special role of anisotropy): measured by our descriptors, there are very few systems with comparable properties $(\mathrm{DOS_F} \cdot b_\mathrm{F}, \bar{v}_\mathrm{F})$. Two of them, both high-pressure phases of $III - V$ semiconductors, have been included in the validation set (cf. Figure 10.1).

In the following, we present a short analysis of the relation between chemical/structural properties and our descriptors, which was inspired by the empirical rules to find new superconductors established by B.Matthias (outlined in chapter 5). To be more specific, based on the assumptions that (a) our scheme predicts superconductors (supported by the low error rate when applied to the validation set), and (b) ICSD consists of a repre-

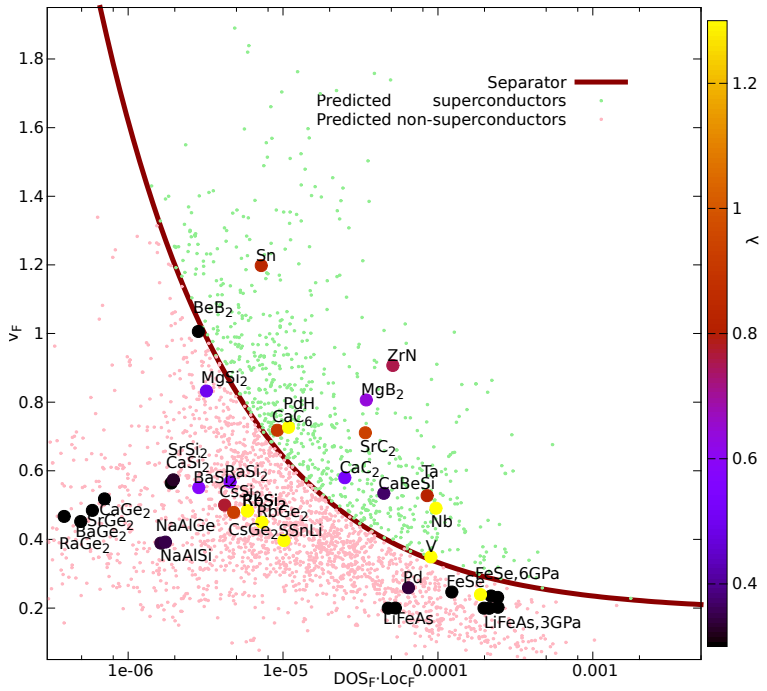


Figure 10.2.: Prediction of superconductivity by (10.2). Dots represent the nonmagnetic metals from our high-throughput search, color corresponds to the prediction. Initial validation set materials (10.1.2) are displayed as circles, color corresponds to λ.

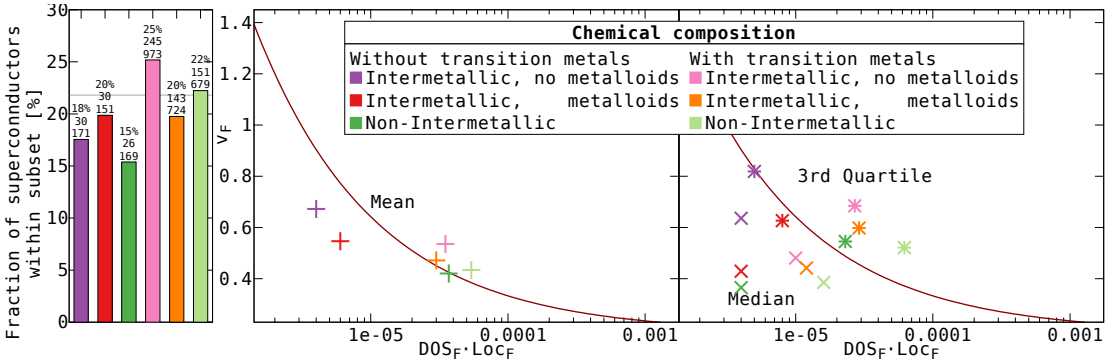


Figure 10.3.: Fraction of materials within each chemical subclasses (left panel) predicted as superconducting, and statistical data of the distributions of the descriptors of superconductivity within the chemical subclasses (center panel: mean, right panel: median and 3rd quartile)

sentative sample of materials, we compare the probability of predicted superconductivity within selected subsets of the materials.

10.3.1. Effect of chemical composition

As a first part of the analysis, we split compounds (*elemental solids* are a special case, and will be presented explicitly in subsection 10.4.2) into subsets according to chemical criteria, on the basis of expected relevance regarding our descriptors of superconductivity; the classes are defined by bonding character, as expected by from simplified chemical considerations, as intermetallic with at least one metalloid (Metalloid-IM), intermetallic without any metalloid element (NoMetalloid-IMs) and non-intermetallic (Not-IM). The first distinction is made due to the increase in covalent bond character in the presence of metalloids. A further separation has been made based on the presence of transition metals (TMs). The relative number of predicted superconductors has been computed in each of the 6 classes and is displayed in the left panel of Figure 10.3.

The presence of TMs in a compound has no effect whatsoever in Metalloid-IMs; 20% of both *sp* and TM systems are predicted to be superconducting. Again independent of the presence of TMs, the fraction of predicted superconductors among NoMetalloid-IMs is 3% larger than in Not-IMs. However, the NoMetalloid-IMs are classified at a significantly higher fraction of 25% as superconductors, compared to the fraction of 18% within the same class and in the absence of TMs.

We continue our analysis by appling statistical methods to the subsets. The distributions of $\mathrm{DOS_F}$,b_F and vF (chapter 8) are heavily tailed and asymmetric, i.e. far from normal, therefore a description by mean and standard deviation is of limited use. Therefore we use the following quantities in statistical analysis:

Mean

Median upper bound of the lowest 50% as a measure for the body

3rd Quartile: lower bound of the highest 25% as a measure for the tail

Moreover, as non-neglegible statistical correlation between b_F and $\mathrm{DOS_F}$ was found (section 8.5), the statistical measures are applied to the product of both quantities.

The mean descriptor values obtained on the different subsets are displayed in the central panel of Figure 10.3, while median and 3rd quartile are displayed in the right panel. The mean $b_\mathrm{F} \cdot \mathrm{DOS_F}$ observed on the *sp* Metalloid- and NoMetalloid-IMs is very close (note the logarithmic axis scale) and about an order of magnitude lower than in all other materials; $\bar{v}_\mathrm{F}$-mean in the two classes is about 20% larger than in the remaining systems. However, all subset means are fairly close to the separator, which implies some difficulty in the explanation of the fractions of predicted superconductors.

At this point, we focus on the other two statistical quantities, the *median* and the *3rd quartile*, displayed in the right panel of Figure 10.3; it is fairly obvious from the almost identical median values that the main body of $b_\mathrm{F} \cdot \mathrm{DOS_F}$ distribution in all *sp* systems is rather similar. The NoMetalloid-IM $\bar{v}_\mathrm{F}$-median is strongly enhanced over the other two *sp* classes, reflecting the low number, or even complete absence of systems in the low-$\bar{v}_\mathrm{F}$ regime, confirming the expected behaviour. Larger median $b_\mathrm{F} \cdot \mathrm{DOS_F}$ can be observed in the three TM classes (again indicating the lower frequency of low-$b_\mathrm{F} \cdot \mathrm{DOS_F}$ systems), while median $\bar{v}_\mathrm{F}$ is very similar to the *sp* Metalloid-IMs and Not-IMs.

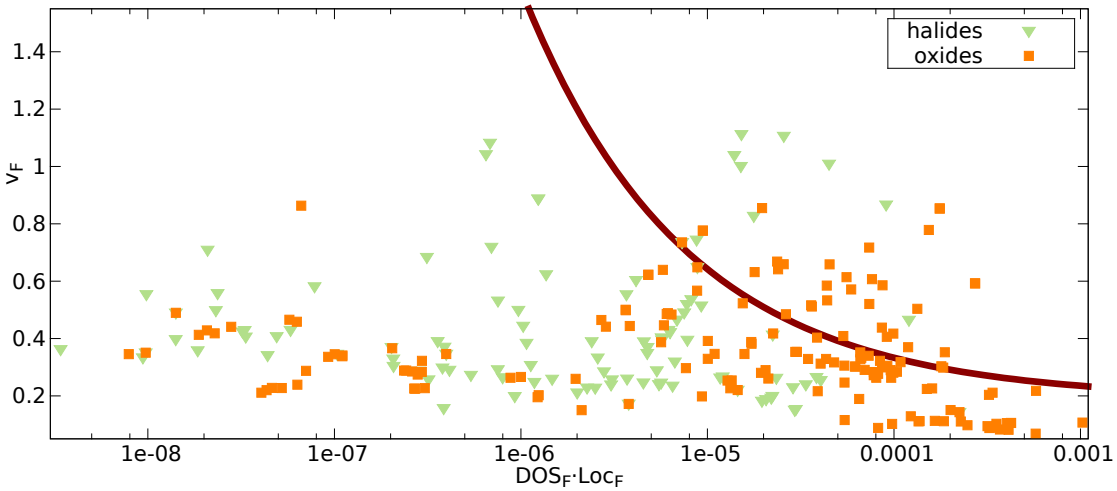


Figure 10.4.: Classification within chemical subclasses of the nonmagnetic metals of the dataset. Points correspond to the subclasses of materials containing oxygen or any halogen. In both subclasses, the S^{dlp}-classification curve removes a significantly higher fraction ($> 82\%$) of materials than on the overall dataset (76%).

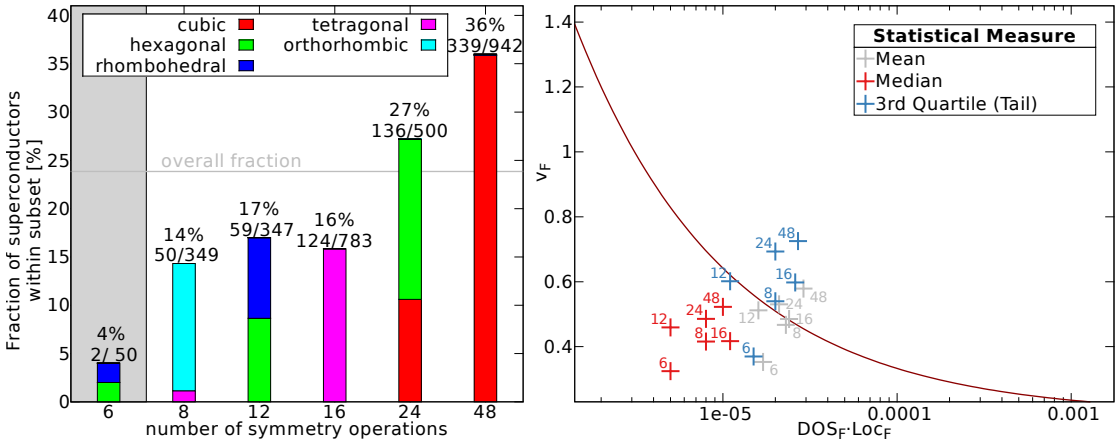
The 3rd quartile finally reveals the reason for the significant differences in SC classification frequency: the difference in the fractions of predicted superconductors is mostly related to the high-value tail of the $b_{\mathrm{F}} \cdot \mathrm{DOS}_{\mathrm{F}}$ distribution, as the 3rd quartile defines the lower bound for the highest 25% of a distribution and thereby provides an estimate of the high-value tail.

Classification properties in Oxides and Halides

The results of the filtering process can be also used to revisit the "rules for finding new superconductors" formulated by B.Matthias, presented in chapter 5. One of the rules reads "Stay away from Oxygen", which reflects within our classification method as a fraction of predicted superconductors of 18% observed on nonmagnetic, metallic oxygen compounds, which lies significantly below the overall fraction of 24%. The lower fraction of such compounds being predicted as conventional superconductors is mainly a result of their accumulation within the range of low Fermi velocities (Figure 10.4), suggesting unreliability of LDA Kohn-Sham DFT in such compounds. Note that this statement does *not* imply a causal connection to Matthias' rule, as $\bar{v}_{\mathrm{F}}$ had been introduced as a descriptor out of methodological reasons.

Another chemical subclasses of nonmagnetic metals experiencing a similarly low fraction of predicted superconductors are materials containing halogens. Such materials tend to exhibit larger $\bar{v}_{\mathrm{F}}$ than the oxygen compounds, but are declassified mostly due to low S^{dlp}.

Conventional superconductors belonging to both classes are rather uncommon, therefore we suggest a systematic study on the 10 halides and 34 oxides predicted by our method.



(a) Fraction of predicted superconductors within the symmetry subclasses; color corresponds to lattice system. Gray background indicates statistical unreliability (low numbers of materials).

(b) Descriptive statistics in the symmetry subclasses, relative to the classification curve. Each subclass is represented by points $(F(\{\bar{v}_F\}), F(\{b_F \cdot \mathrm{DOS}_F\}))$, where F is a statistical measure on the subset.

Figure 10.5.: Influence of space group symmetry on the fraction of predicted superconductors: space groups are summarized by the number of symmetry operations.

10.3.2. Influence of symmetry on the fraction of predicted superconductors

A further statistical observation of classification probability within subclasses of the non-magnetic metals of the dataset is again inspired by one of Matthias' rules, in which he suggests that high, and especially cubic symmetry has a positive effect on superconductivity.

Within our method, this effect is observable as a statistical correlation between fraction of predicted superconductors and symmetry: Figure 10.5 displays the fraction of materials predicted as superconductors for materials of different symmetry classes. In our representation, each of the classes summarizes space groups by their size, which corresponds to the number of symmetry operations applicable in a material. Each bar is subdivided by the *lattice system* in order to provide further information on the lattice and to indicate symmetries broken by the basis, i.e. the atoms in the unit cell.

As a first remark, there are very few low-symmetry materials available within the dataset, therefore statistics should be considered unreliable within the area highlighted with a gray background.

A trend consistent with Matthias' observations is clearly observable: materials with high, especially complete cubic symmetry exhibit a significantly higher fraction of predicted superconductors than materials with lower symmetry (although one should be careful about the statistically badly sampled classes with 6 or less symmetry operations). The difference is striking: 36% of the systems with full cubic symmetry (48 symmetry operations) are predicted superconductors, while only 14% of the systems

with 8 symmetry operations are considered superconducting. This range exceeds the one found among chemical subclasses (subsection 10.3.1), and in fact a secondary analysis taking into account the symmetry in each of the chemical classes shows exactly the same trend. Furthermore, also the fact that a significant number of elemental solids (subsection 10.4.2), a part of them high-pressure phases, exhibit cubic symmetry does only have minor impact on the high-symmetry fraction of predicted superconductors (excluding them would lower the fraction by about 3%).

It is therefore justified to conclude that higher symmetry is genuinely correlated with a higher probability of superconductivity, in agreement with Matthias' rule.

In the right panel of the figure, statistical data for each of the subclasses is displayed; as can be observed on all statistical quantities within the subclasses, the main component of symmetry-related variation is orthogonal to the classification curve. As in the case of the chemical subclasses, the largest spread can be observed in the 3rd quartile (blue points), indicating that it is especially the high-value tail of the descriptor distributions which differs among the symmetry subsets. The different mean values in the subsets (displayed in gray) can be explained by the tail part of the distributions, too; the variation in the main body, as described by the median (displayed in red), on the other hand, is far too weak to explain the mean value variation.

10.3.3. Summary

In this section, classification properties of our method (section 10.2) have been evaluated. First of all, in conjunction with the filtering steps described earlier in section 8.1, the method allows to exclude more than 90% of all materials as (phononic) superconductors, with an expected prediction accuracy of 90% (subsection 10.2.5).

A subset of the empirical *Matthias rules* (chapter 5) corresponds to the pre-filtering process (8.1), such as the exclusion of magnets and insulators. The analysis within the present section shows that also the remaining rules are recovered in the form of statistical trends: the presence of oxygen significantly lowers the probability of superconductivity, while higher symmetry correlates with a higher probability to find a superconductor.

Furthermore, an analysis of different chemical classes provides suggestions for a future search for superconductors, exploring chemical compound space beyond the boundaries of the inorganic crystal structure database (ICSD): the highest fraction of superconductors is predicted among systems which (a) are composed solely of metal elements and (b) contain at least one transition metal.

Having now established classification properties of our method, the following section will focus on the properties of the predicted superconductors.

10.4. Predicted superconductors

The previous sections introduced our method to predict superconductors on the basis of computationally cheap first-principles calculations, and a semi-empirical classification scheme. Basic properties of the scheme were evaluated statistically by application to a set of materials from a database.

class	elemental	binary	ternary	quaternary
N_{sc}	125	464	157	4
p_{sc}	45.0%	26.7%	14.2%	17.4%

Figure 10.6.: Distribution of predicted superconductors among elemental, binary ternary and quaternary solids. p_{sc} is the probability of a nonmagnetic metal, belonging to the given class, to be predicted as superconducting. The overall probability of a NM metal from our dataset being predicted superconducting is 23.8%.

The central result of that application is a set of 750 materials, which were predicted to be superconductors. In this section, we present an overview over this set, also introducing single examples, families of materials and trends.

Detailed data on all compounds, such as chemical composition and crystal structure can be found in Appendix B, which may directly serve as an inspiration for future in-depth studies of selected materials.

10.4.1. Distribution among elemental, binary, ternary and quaternary solids

For the following steps of the description of our predicted superconductors (PSCs), it is convenient to separate the set of materials by the *number of different elements in the chemical composition* into elemental, binary, ternary and quaternary solids (Figure 10.6). First of all, quaternary solids do provide two orders of magnitude less materials than any of the other three classes, a fact that is mainly related to the maximal primitive cell size of 7 atoms within the dataset.

The values p_{sc} denoted below the segments of Figure 10.6 are the frequency of PSCs within the corresponding subset of the nonmagnetic metals. The fraction among elemental (45%) and binary (27%) solids is significantly higher than the overall one (24%), while a significantly lower fraction of materials composed of more than 2 different chemical elements qualifies as superconductors: in the original dataset, binary and ternary systems were contributing both very similar numbers of materials, each about 47%, while the ternary systems contribute only 21% of the PSCs.

The difference in the fractions between elemental, binary, ternary and quaternary solids can be explained by the relation between prediction and symmetry already discussed in subsection 10.3.2: high values of $\mathrm{DOS}_F \cdot b_F$ and $\bar{v}_F$ are statistically correlated to the number of symmetry operations; within our dataset, there is an upper bound $N_{at} \leq 7$ on the cell size, which seriously limits the possible symmetry of systems composed of a larger number of atoms (postponing the role of *stoichiometry* to subsection 10.4.5). This physically intuitive argument was confirmed by a corresponding statistical analysis.

In the following sections, we will first discuss the elemental solid PSCs, introduce general chemical features of the compounds and review the binary, ternary and quaternary subclasses.

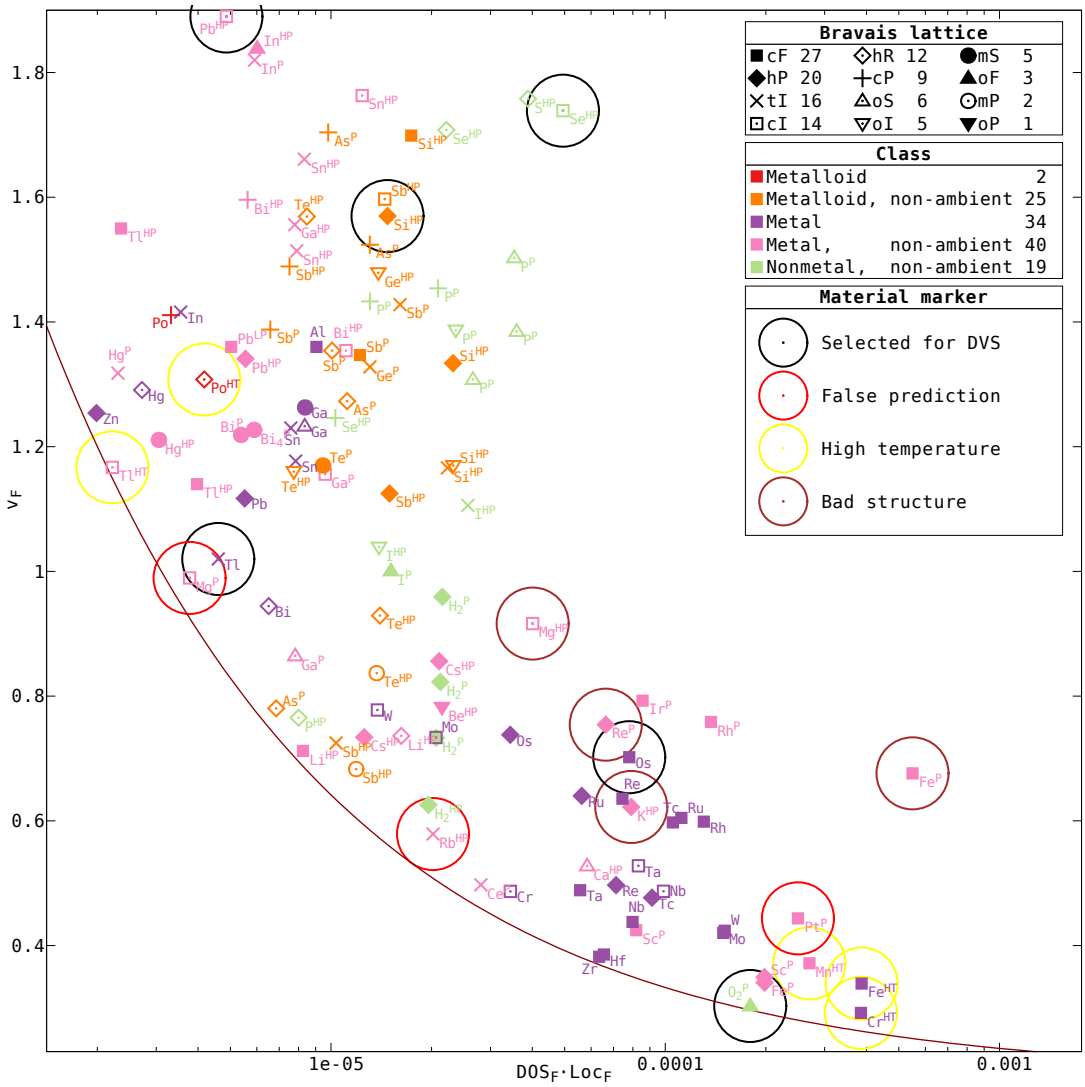


Figure 10.7.: Elemental solids classified as superconductors by our method, and their position within the classification diagram. The shape of each point refers to the Bravais lattice, while the color refers to the class (metal, metalloid, nonmetal) and pressure/temperature of the phase, if such data is available. Superscript (H)P indicates (high) pressure phase. Numbers in the legend refer to the respective number of elemental solid PSCs.

10.4.2. Elemental solids

Structural phases of the periodic elements at various pressures and temperatures have been subject to very detailed research in the past and are therefore strongly represented within ICSD. As a consequence, among the 750 predicted superconductors (PSCs), there are 125 distinct elemental solids, which are different phases of 47 chemical elements. The elemetal PSCs are displayed as a classification diagram in Figure 10.7.

The majority of those structures (84, corresponding to 67%) are elemental phases under pressure (indicated by a P or HP superscript),the largest subgroup consisting of metalloids and nonmetals, such as hydrogen and the relevant members of the *p*-block. The shape of each point corresponds to the Bravais lattice of the elemental solid; as can be read from the numbers provided in the legend box, the most PSCs are found having either face-centered cubic (cF) and hexagonal (hP) structure. The color of each point refers to the chemical class of the element (hue) and pressure (saturation) of the structure. Trends can be fairly easily read from the colors, listed in the order of $\mathrm{DOS_F}{\cdot}b_\mathrm{F}$:

post-transition metals are found in the region of low $\mathrm{DOS_F} \cdot b_\mathrm{F}$ at high $\bar{v}_\mathrm{F}$, together with the group-12 transition metals Zn and Hg and Po at ambient pressure,

metalloids under pressure follow, with a wider range of $\bar{v}_\mathrm{F}$,

nonmetals under pressure are the next class, again enlarging the $\bar{v}_\mathrm{F}$-range downwards,

transition metals with and without applied pressure are found in the final region of $\mathrm{DOS_F} \cdot b_\mathrm{F}$, while restricted to lower $\bar{v}_\mathrm{F}$; among them, also a low number of high-pressure alkaline- and alkaline earth metals can be found.

In the following discussion, we will use the data collected in [167] as a reference for the superconducting properties of elemental solids.

Owing to infrequently occuring numerical issues, which were discussed in in section 5.3.2, the Fermi bond localization b_F of 4 high-pressure phases of Pb, Mg, Rb and Pt is overestimated, resulting in their classification as superconductors (marked by a red surrounding circle in Figure 10.7).

A set of other elemental phases, marked by a brown surrounding circles, corresponds to apperently faulty entries in ICSD, as the computed pressure of the report structures, does by far exceed the pressures reachable by modern diamond-anvil cells, including a large error margin due to the LDA exchange-correlation approximation.

A third set of apparently false predictions, marked by yellow surrounding circles, are actually high-temperature structures, such as the nonmagnetic Fe^{γ} with face-centered cubic (cF) structure, observed at about 1200 K. As a sidenote, superconductivity has actually been observed [168] in the one remaining iron high-pressure hexagonally closest packed structure.

The apparently false predictions of superconductivity in crystal structures unrealistic at low temperature illustrate the fundamental fact that our method has been designed to predict *superconductivity* at low computational cost; *stability* of a structure at a

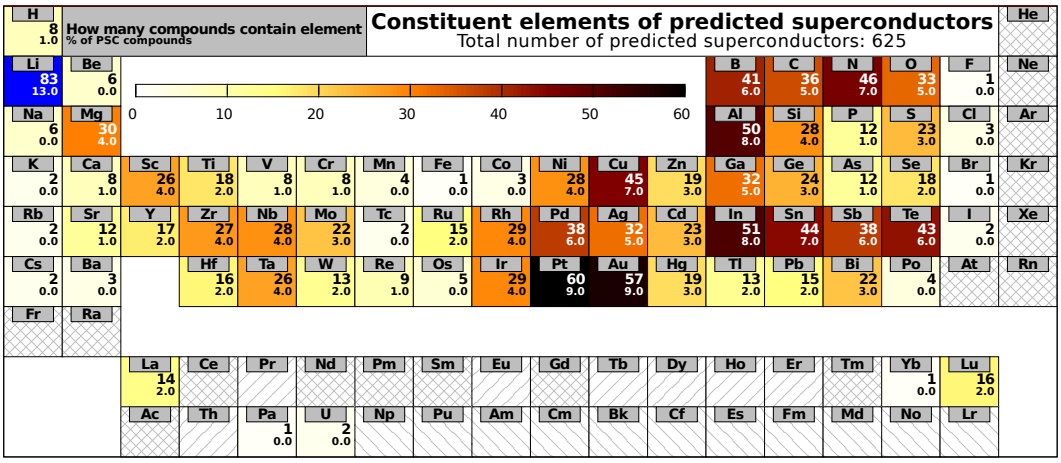


Figure 10.8.: Constituent elements of predicted superconductors. Color (and number in first row) corresponds to the absolute number of materials containing the given element, and as lithium is occuring 40% more often than any other element, it has been labeled in blue, while adjusting the colorscale to the remaining systems. A background pattern is used for each element not contained in any of the predicted superconductors.

given temperature and pressure has to be determined by other methods (in our high-throughput search, structural relaxation by damped dynamics is the first task performed on selected candidate materials).

These exceptions set aside, the classification by our method is in agreement with the data provided in [167]. Predicted superconductivity in Po (cP lattice) has been confirmed by first-principles calculations [169]; however, due to the difficulties (Po is a short half-life α emitter, isotope with largest half-life of 125 y is ^{209}Po) for handling actual samples of Po, (a) experimental references could not be found and (b) the prediction is of little practical relevance.

As phonon calculations of elemental phases are in general computationally cheap, a subset of materials has been chosen for inclusion in DVS (marked by black surrounding circles). As can be read from Figure 10.1, predictions and computed λ agree in all cases except for (a) the high-pressure Pb phase at the upper limit of $\bar{v}_\mathrm{F}$range due to the aforementioned reason and (b) the high-pressure O_2 phase on top of the curve in the high-DOS$_\mathrm{F} \cdot b_\mathrm{F}$regime, which during structural relaxation obtains a $\bar{v}_\mathrm{F}$below the lower bound, implying that the originally sampled structure was non-equilibrium.

After having illustrated the predictions of elemental solids, we will now turn towards the compounds predicted to be superconducting.

10.4.3. Constituent elements of the predicted superconducting compounds

A coarse overview on the chemical composition of the 625 compounds predicted as conventional superconductors (PSCs) is presented in Figure 10.8, in the form of the periodic table of elements, where the number of PSCs containing a given element is both

given as a an absolute number and determines the color of the element's box.

First of all, the effects of the filtering descriptors (section 8.1) can be observed in the absence of any noble gas compound, as all of them are insulators; this result is expected from basic chemical considerations. Moreover, the majority of block-center $f-$ and $3d$ transition metal compounds had been filtered due to magnetism (cf. Figure 8.2), which was also an expected result; therefore, combined with the classification described in the present chapter, only 3 lanthanides and 2 actinides occur in PSCs.

Lithium is the element found in more predicted superconductors than any other one, starting a diagonal trend of frequently found elements which extends into the d-block: Li, Mg, Sc, Zr and Ta are all found more frequently within predicted superconductors than the original distribution within the dataset would suggest (cf. Figure 7.7); in principle, due to its chemical similarity to Ta, Nb could be seen as part of the same trend.

It is surprising to find a significant number of gold and platinum systems among the predicted superconductors, surprising because few such superconductors are known in practice. We therefore choose to compute the superconducting properties of one such system, $AuTe_2$ under pressure, in order to assess if these predictions could be an artefact, i.e. a false-positive, of our classification scheme. The material was found both to be structurally stable and superconducting, which confirms the validity of the proposed classification scheme, and further research showed experimental evidence of superconductivity in this system [160]. A single confirming example, of course, does not suffice for a statistical conclusion; therefore future research should be performed on this class.

Copper, part of 45 PSC compounds, is the next frequently found transition metal; while most famously known for its role in the *cuprate* high-temperature superconductors, all such CuO_2 layered cuprates have been discarded on the basis of a too low $\bar{v}_F$, characterizing them (correctly) as strongly correlated systems, which could not be correctly described within LDA. Out of the remaining 45 Cu-compounds, the $CuAl(S,Se,Te)_2$ compounds unter pressure with Cu_2MnAl structure (Figure B.7) and $CuAl_2$ (structure type in Figure B.2) have been selected as members of DVS, all of which have been confirmed as superconductors.

The $5p$ metals and metalloids contribute on a similarly high level (each around 7%) as Al, and slightly higher at the $2p$ metalloids and nonmetals, which are found in about 6% of the predicted superconductors.

10.4.4. Influence of chemical composition and pressure on the distribution of predicted superconductors

After introducing an overview about the abduance of each periodic element among the predicted superconducting (PSC) compounds, we will now review the position of materials on the more coarsely defined chemical classes we introduced in subsection 10.3.1; however, within this section, instead of the statistical quantities, the actual distributions are reviewed.

Keeping the 125 elemental phases aside, which were presented separately in subsection 10.4.2, the chemical composition of the vast majority of PSCs does contain at least

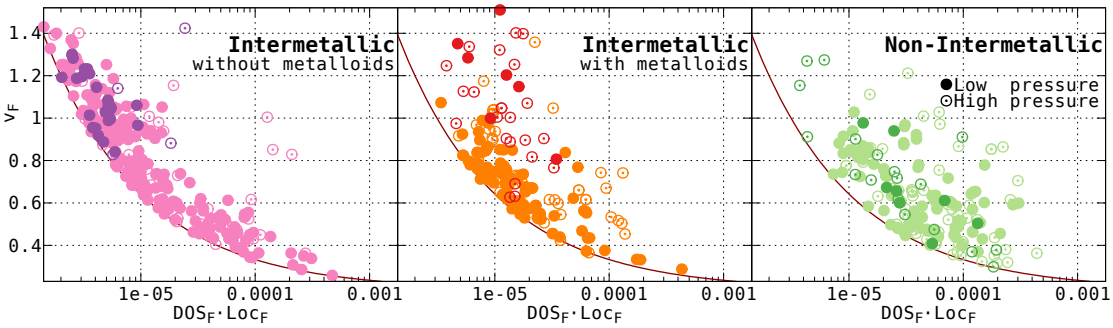


Figure 10.9.: Distribution of predicted superconductors with respect to the classification curve for intermetallic and non-intermetallic compounds. The former have been further divided by absence or presence of metalloids (B, Si, Ge, As, Sb and Te). Filled circles correspond to low- or ambient pressure phases; empty circles to pressurized phases. Darker color is used in the absence of transition metals.

one metal (611) or, in absence of the latter, metalloid (13 materials) atom within their chemical composition; only one example not containing any metal or metalloid atom, a rocksalt-structure high-pressure phase of SeS (cf. Figure B.1) exists among all PSCs, and was found to be unstable.

An important fact about the predicted superconducting compounds is the presence of of at least one transition metal atom in 539 of the 625 compounds (86%). Compared to the *sp* nonmagnetic metals, of which 17.5% are classified as superconductors by our method, transition metal compounds are also classified with higher probability (22.7%) as superconductors. In this section, while discussing the influence of chemical composition on the position of a material, the presence of transition metals is found to lead to higher $\mathrm{DOS_F}$ and b_F among the predicted superconductors. As a sidenote (cf. subsection 10.3.1): the mean of both quantities (over the whole set of nonmagnetic metals) is about 70% higher in TM compounds compared to *sp* ones, while the $\bar{v}_\mathrm{F}$ mean is about 20% lower.

The distance of a point $(\mathrm{DOS_F} \cdot b_\mathrm{F}, \bar{v}_\mathrm{F})$ representing a material from the separator corresponds to the reliability of our prediction. Moreover, empirical data obtained on the descriptor validation set does support, up to the unknown amplitude factor (discussed in section 10.2.4), the relation between the strength of the electron-phonon interaction λ and $(\mathrm{DOS_F}, b_\mathrm{F})$.

On the other hand, materials with larger $\bar{v}_\mathrm{F} \gg 1.0$ are underrepresented within DVS, therefore classification is also expected to be less reliable in this region.

Due to all three reasons, materials found to the north-east of the separator and within well-characterized $\bar{v}_\mathrm{F}$-range, are predicted to be superconductors with increased confidence. The first question to be answered is: what influence does the chemical composition have on the position of the corresponding point? A second question which occurred after discovering that a good number of high-pressure phases were predicted to be superconducting: how is the relation between pressure and the position of the point?

While the sheer number of materials is too large for an in-depth discussion within this

section (chemical composition and crystal structure of PSC compounds is presented in Appendix B), trends can be demonstrated within more coarsely cut chemical classes:

In Figure 10.9, data for predicted superconductors belonging to three classes is presented: Intermetallic compounds, further subdivided into those without and those with metalloids as constituent elements, and non-intermetallic compounds. Coarse information about pressure is represented by the shape of each point: given that high pressure is indicated by either information found within ICSD or within our DFT calculations, the corresponding point is drawn as an empty circle; all other cases are indicated by filled circles.

Pressure does have a strong influence on our descriptors: in all three cases, the upper right of the figure is dominated by high-pressure phases, having significant $\bar{v}_\mathrm{F}$ and $\mathrm{DOS}_\mathrm{F} \cdot b_\mathrm{F}$ simultanously.

Comparison of the three subfigures clearly shows that compounds solely consisting of metals (the first group) are predominantly found very close to the separator and at higher $\bar{v}_\mathrm{F}$; therefore, they represent less promising candidates than those which contain metalloids or nonmetals.

The introduction of metalloid atoms leads to a significant shift towards materials lying within the medium $\bar{v}_\mathrm{F}$ range, and large $\mathrm{DOS}_\mathrm{F} \cdot b_\mathrm{F}$ ($\mathsf{MgB_2}$ does belong to this class, and is represented by the dark-red filled point about the center of the figure). Addition of nonmetals (including halogens) to the chemical composition leads to a further increase of materials found in the desired region. Both results reflect the hypothesis of superconductivity being enhanced by the presence of localized Fermi bonds, which we outlined in section 5.3: the formation of bonds with covalent character at ambient pressure occurs in the presence of metalloids and nonmetals with increased frequency, while the metallic constituents provide the doping necessary to achieve partial filling of these localized bond states, the necessary condition for phonon-induced transitions among them.

As a further detail, in Figure 10.9 the brightness of the respective color corresponds to the presence (dark color) or absence (brighter color) of transition metals (TMs) within the compound. As mentioned earlier, the vast majority of predicted superconductors does contain TMs. While the effect of TMs is present within all three panels, it can be most easily observed in the left panel, i.e. in the absence of metalloid and nonmetal atoms: stronger localization at low pressure can only be observed in the presence of TMs; their absence results in lower localization and DOS_F, but increased Fermi velocity. This observation suggests that the d and f states provide channels for Fermi charge localization in the absence of metalloids and nonmetals. Furthermore, as the position of TM-, metalloid- and nonmetal-free materials lies close to the upper edge of the curve (where the parametrization of the separator could not be properly refined due to the lack of training examples), a notable rate of false positives among them is expected.

10.4.5. Stoichiometry and compounds

The 464 binary compounds represent the largest group of predicted superconductors; the distribution among different stoichiometries is presented in Figure 10.10a. Ternary

Stoichiometry	AB	AB_2	AB_3	other
N_{sc}	259	157	41	7
p_{sc}	34.1%	22.8%	23.6%	6.3%

(a) Binary PSCs; for comarison: 26.7% of the binary nonmagnetic metals are PSCs.

Stoichiometry	ABC_2	ABC	ABC_3	AB_2C_2
N_{sc}	100	23	19	15
p_{sc}	39.7%	6.7%	13.7%	5.1%

(b) Ternary PSCs; within this class, 14.2% are PSCs.

Figure 10.10.: Distribution among stoichiometries of predicted superconductors

compounds are with 157 members the second largest group, and their distribution among stoichiometries is presented in Figure 10.10b.

In both panels, the number of predicted superconductors having the given stoichiometry are printed below the bar segments, alongside the corresponding fraction within the subset. Starting with 40% among $\mathsf{ABC_2}$ compounds, the probablitity of superconductivity gets succesively lower in AB (34%), $\mathsf{AB_3}$ (24%), $\mathsf{AB_2}$ (23%), $\mathsf{ABC_3}$ (14%) and the remaining ones, which are predicted with around 5% probability to be superconducting, which, as in subsection 10.4.1 is correlated with the symmetry of materials having the given stoichiometry, when the maximal cell size is limited to 7 atoms.

Conclusion

This goal of this work has been the development of strategies for the application of high-throughput methods to search for better superconductors. The main difficulty is that while ab-initio, predictive theories for superconductivity exist [36, 37, 47, 170, 171] their computational cost is too high to be applied in this context.

A first attempt has been to adopt machine learning methods to skip computationally expensive calculations completely. Our results, although promising, show that the accuracy of these prediction methods still leads to an error bar that is too high to provide reliable predictions in the subtle field of superconductivity.

The strategy must therefore be the formulation of *descriptors of superconductivity*, quantities which allow to identify superconductors at low computational cost (in the order of a Kohn Sham electronic structure calculation). In this work we have constructed and discussed a set of such descriptors for phonon-driven superconductivity. Some of them are trivial exclusion criteria (such as magnetism), while others are less obvious. The most important one we have constructed in this work is the 'so called' Fermi bond localization (section 5.3), that correlates well with the strength of the electron phonon coupling.

This first generation of descriptors has been computed on a set of about 8000 materials from the ICSD, and validated on a set of about 80 materials, leading to an accuracy of about 80%. This accuracy is already quite high and therefore the method can be considered ready for predictive approaches. As future development it will be important to test beyond the domain of existing materials and move to new ones, i.e. to combine its predictive power with modern in silico synthesis methods [134–138].

Looking forward to this future application, in light of statistical analysis performed on the ICSD data, in this work we have also investigated how the existing experimental knowledge can be practically used for the prediction of new structures. This has lead to the formulation of a scale of similarity between chemical elements (in line with the original idea of Pettifor [140, 148] in the 80s).

We believe that this work poses the basis for future developements and for improving the role of theoretical methods in the search of new, better superconductors.

Part III.

Appendix

A. Computational implementation

A.1. Implementation of the Bader Fermi Bond Localization

A.1.1. Grid approximation of Bader atomic volumes

We employ a grid method [172] for the Bader analysis of ρ, where the analytical description of the surface is substituted by an assignment of grid points to the Bader atomic volume, i.e. the result of the analysis is

$$A(\boldsymbol{r}_i) \quad \text{(atom index within the unit cell) for each grid point } \boldsymbol{r}_i. \tag{A.1}$$

In cases when an atom has a periodic replica of itself as nearest neighbour, this method would not succeed (in the case of a monatomic crystal, no surface at all would be detected, as all grid points are assigned to the *same* atom index). We perform the analysis on a $2 \times 2 \times 2$ supercell in order to circumvent this limitation.

Furthermore, the method relies on the presence of *maxima* in the charge density close to the nuclei, an assumption that is not necessarily fulfilled when the nuclear Coulomb potential (and the electronic core states) is replaced by a pseudopotential, especially in the case of alcaline and alcaline earth metals. In our calculations, we reconstruct the core charge either by the data present for non-linear core correction (NLCC), or in its absence, by narrow Gaussians ($\sigma = \frac{1}{3} r_{\text{covalent}}$) centered at the nuclei, which are normalized to reconstruct $\frac{1}{3}$ of the pseudized core charge (both parameters have been determined empirically by analysis of the NLCC charge represented on the grids commonly found in our calculations). The changes in the total charge density, and more importantly its gradient, are neglegible in the surface regions, whigh in turn means that it has neglegible effect on the outcome of the grid-based bader analysis.

A.1.2. Approximation of the Bader surface on a grid

However, our goal is to find an approximation for the Bader surface $\mathcal{B}$ on the discrete grid in order to perform our bond localization analysis (5.10). A single grid point is associated to a finite volume, therefore we approximate the surface integral by a volume integral

$$b_{\text{F}} \approx \lim_{\varepsilon \to 0} \int_{\mathcal{B}_\varepsilon} |\boldsymbol{\nabla} \rho(\boldsymbol{r})| \mathrm{d}V / \int_{\mathcal{B}_\varepsilon} \mathrm{d}V$$

on a small shell $\mathcal{B}_\varepsilon$ around the bader surface. We compute b_F by discretizing the integral via the mid-point rule, which reduces the expression to the arithmetic mean

$$b_\mathrm{F} \approx \frac{1}{n_b} \sum_{i=1}^{n_b} |\boldsymbol{\nabla}\rho(\boldsymbol{b}_i)| \tag{A.2}$$

over all n_b grid points $\boldsymbol{b}_i$ within $\mathcal{B}_\varepsilon$. In our calculations, $\{\boldsymbol{b}_i\}$ is determined from $A(\boldsymbol{r}_i)$ (A.1) by the result of an adaptive-mask finite-difference Laplacian convolution $\tilde{L}(\boldsymbol{r}_i)$.

A.2. High-throughput search with Quantum Espresso

We employ Quantum Espresso [173] (QE), a comprehensive electronic structure code, in our high-throughput search for superconductors, in which we did also implement the bond-localization method described in section A.1.

In this section, we report relevant information on Quantum Espresso. Furthermore, we describe mechanisms developed by us in order to improve the reliability of the computations, including automatic error recovery.

A.2.1. Input creation

The crystal structure and computational parameters, such as the $\boldsymbol{k}$-point sampling, the choice of pseudopotentials, the size of the plane-wave basis and the real-space grid point densities, need to be specified to QE by means of a simple text file, in the following called "QE input". Manually writing such text files for thousands of materials is impractical, therefore the creation of QE input has been automatized in our work.

Crystal structure from ICSD

As mentioned in section 7.1, ICSD describes a crystal structure by spacegroup and Wyckoff positions of the atoms, accompanied by lattice parameters (such as lenghts and angles), which can be exported in the format of a *Crystallographic Information File* (CIF) by utilities provided by the manufacturer.

Quantum Espresso, on the other hand, requires a description by a complete set of coordinates of atoms within the primitive cell, and the type of the Bravais lattice with the associated parameters.

The necessary conversion between the two conventions is handled by the program `cif2cell` [174], which performs the conversion between ICSD/CIF and the input format used by some other electronic structure codes such as elk [175], and was extended within the context of our work to include support for QE.

Brillouin zone sampling

In order to obtain accurate results from Kohn-Sham DFT calculations, the grid of $\boldsymbol{k}$-points employed for sampling the first Brillouin zone (1BZ) is required to be adapted to

Material	Monkhorst-Pack grid	Material	Monkhorst-Pack grid
MgB_2	$24 \times 24 \times 18$	KO_2	$22 \times 18 \times 18$
CaBeSi	$20 \times 20 \times 16$	Al	$28 \times 28 \times 28$
Nb_3Sn	$12 \times 12 \times 12$	Pb	$22 \times 22 \times 22$

Table A.1.: Examples for Brillouin zone samplings generated by our method, which are used by our calculations in conjunction with a fixed Gaussian smearing of 0.02 Ry

each individual crystal structure. In our calculations, the Monkhorst-Pack-method [56] is used for sampling the 1BZ by the means of an unshifted (i.e. including the $\mathbf{\Gamma}$ point), regular grid, parametrized by the number of sampling points N_{k_1}, N_{k_2}, N_{k_3} along each of the reciprocal axis $\boldsymbol{b}_1$, $\boldsymbol{b}_2$ and $\boldsymbol{b}_3$.

In conventional calculations, these three parameters are determined by means of *convergence tests*: for a fixed smearing width σ, the convergence of a quantity under increasing N_{k_i} is observed, giving an estimate for the accuracy achieved by each grid setup. Based on this data, a tradeoff is then made between required computation time and achivable accuracy.

In the context of our high-throughput search for superconductor the computional resources required for such tests would largely exceed the available ones. We note, however, that at this stage a higher accuracy is of lower priority.

A simple estimator for the 1BZ sampling grids has been employed in this work, based on the assumption that the number of sample points along one reciprocal axis should be proportional to the length of that particular axis:

$$N_{k_i} = \lceil p_k \cdot |\boldsymbol{b}_i| \rceil_2 \,,$$

i.e. in our case, the number of sampling points along each axis is taken as the closest *even* integer, larger or equal then the length of that axis times a proportionality factor p_k. The value of the factor $p_k = 18.37$ has been determined empirically by convergence tests on a small set of materials, taking also the number of states close to E_F into account, as the present implementation of the bond localization does not include an interpolation scheme for Kohn-Sham wave functions and the resulting densities. We perform all our calculations with a Gaussian smearing of 0.02 Ry, which we employed also for this convergence test. A few examples of grids generated by the described method are reported in Table A.1.

Exchange-correlation functionals and magnetism

Our calculations employ spin-resolved local density approximation (LSDA) as exchange-correlation (xc) functional, mostly due to computational constraints. For one, evaluation of an LSDA functional involves less computational effort than some of the more complex xc functionals, including the ones from the generelized gradient approximation (GGA) family; furthermore, GGA functionals tend to require denser direct-space grids for an

accurate evaluation of the charge density gradients, leading to a further increase of the computational complexity.

Collinear LSDA is used whenever d and f transition metals are present in a system; the local spin polarization on each of the transition metal used for the initial symmetry breaking is initialized to a ferromagnetic configuration. While the result of such a setup may only lead to ferromagnetic solution, breaking symmetries in order to allow also antiferromagnetic solutions, which may require the expansion of the original unit cell into *supercells*, is deemed to be too (computationally) expensive in the context of this work.

In the absence of d and f transition metals, the conventional local density approximation (LDA) exchange-correlation functional is used, due to a neglectable number of known pure sp systems exhibiting a spin-polarized ground state, and the resulting significantly lower computational complexity.

Pseudopotentials, basis size and direct-space grids

Quantum Espresso (QE) represents Kohn-Sham states expanded in a plane-wave basis set. In order to achieve convergence with a finite-sized basis set, the nuclear coulomb potentials are replaced by pseudopotentials much softer than the $Z/|r|$ coulomb potential, which contain a non-local, i.e. l-dependent component [176, p. 22ff].

We chose the set of norm-conserving LSDA pseudopotentials generated at Fritz-Haber-Institute (FHI) [177] for our high-throughput search for superconductors, as this is a consistent set covering most of the periodic table. Convergence of the bond localization with the size of the basis (parametrized as a kinetic energy cutoff e_{cut}) has been tested on elementary compounds for all pseudopotentials; a relative error of less than 20% is reached for $e_{\mathrm{cut}} = 100\,\mathrm{Ry}$ in the majority of the elementary compounds. Exceptions are located in the f-block of the periodic table, where nevertheless the relative errors lie below 30%. This value, which indirectly also determines the size of the representing the density, has then been chosen for all of our high-throughput calculations.

Parallelization options

Another related topic is the automatic setup of *parallelization* options: Quantum Espresso can perform a simulation on many processors simultanously, via the message passing interface (MPI). The necessity of such a parallelization mainly depends on the computational complexity of the problem, determined by the number of bands (i.e. Kohn-Sham states) and the dimensions of both real and reciprocal space grids. Throughput, measured as the $1/T_{\mathrm{wall}}$[1], does not scale linearly with the number of processes N_{proc} working on the solution of the problem, i.e. running in parallel on twice the number of processors does take *longer* than half of the original time. Involving too many, relative to the size of the problem, processors in its solution may actually increase T_{wall} due to the time processes spend communicating with each other. Most high-performance

[1] T_{wall} is the so-called wall-clock-time, i.e. the real time consumed during the solution of the Kohn-Sham equations

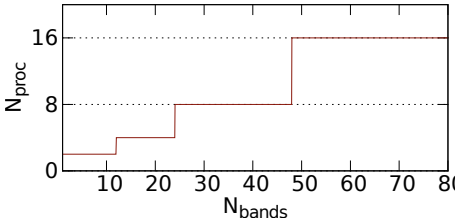


Figure A.1.: Automatically determined parellelization for a given number of Kohn-Sham states N_{bands}

computing facilities, such as the one used in this work, are organized as clusters, where many independent nodes, each containing a comparitively small ($N_{\text{CPU}}^{\text{node}} < 30$) number of processors, are interconnected by a network. Two sources of machine-related errors arise from such a setup: failure of a single node, and failures of the network. One can easily see that the total probability of failure of a calculation due to machine-related problems is dominated by two terms

$$p_{\text{job}}^{\text{fail}} = p_{\text{node}}^{\text{fail}} N_{\text{nodes}} + p_{\text{connection}}^{\text{fail}} N_{\text{nodes}}(N_{\text{nodes}} - 1)$$

with

$$N_{\text{nodes}} \approx \left\lceil \frac{N_{\text{proc}}}{N_{\text{CPU}}^{\text{node}}} \right\rceil ,$$

which emphasizes again the desirability of a minimal number of processes.

Moreover, the time spent on a single job does not play a major role in a high-throughput search (with the exception of runtime limits imposed by the policy of the computing facility; most facilities limit the runtime of an individual computation job to a few hours or days). Keeping this in mind, we define an automatic setup empirically:

$$N_{\text{proc}} = \max(2^{\lceil \log_2(N_{\text{bands}}/6) \rceil}, 2),$$

where the computational complexity is estimated from the number of Kohn-Sham states N_{bands} and a power-of-two number of processes is enforced, with a minimum of 2 processes per job due to a restriction on single-process jobs imposed by the computation facility. Figure A.1 illustrates the resulting dependence of N_{proc} on N_{bands}.

We note that the N_{proc} determined by our estimator is not optimal for QE's default parallelization method, which suggests N_{proc} to be a divisor of the real-space-grids' third dimension. Further development, taking into account also this optimization, is left open for future work. However, in the context of the comparatively small systems, on which we performed the high-throughput search at this stage, we consider our method as sufficient.

A.2.2. Error detection and recovery: the Job Supervision Framework (JSF)

In a high-throughput search, thousands of calculations need to be performed. *Manually* detecting and solving errors in such a large number of computational jobs is fundamen-

tally impossible, both due to limits of human time, and the fact that such a repetitive and therefore subjectively boring task implies an increase in *human* error.

In this section, we present our methods developed for this work, which allow for a high level of automation of the tasks of error detection and error recovery. We call this part of the work the *Job Supervision Framework* (JSF). While the problems mentioned in the following may seem abstract and rare at first, the short analysis on our dataset presented in subsection A.2.3 demonstrates the relevance of these issues.

Origins of errors

The errors observed in the preparation of our high-throughput search can be traced back to three main origins.

Numerical problems may arise within the Quantum Espresso suite of electronic structure codes, such as failing convergence during the self-consistent solution of the Kohn-Sham equations. Furthermore, this class includes problems within the grid-based Bader volume partitioning scheme, which we employ in the in the context of the Fermi bond localization quantification.

Machine problems are a second possible origin of failure during computation jobs. Under this label, we summarize all failures unrelated to numerical problems, happening on a lower level than the numerical computation, such as crashes of the high-performance compute nodes, the network interconnect used by parallel (MPI) jobs or the file system where large, temporary files such as potentials and wave functions are stored.

Resource request problems are related to underestimated resource requests at submission time, as computation jobs may be aborted by the cluster's *batch system* due to excessive (beyond the requested limits) resource usage, such as memory or computation time.

Detection and solution of numerical problems

Numerical problems are comparatively easy to identify, as Quantum Espresso writes the respective information directly in its status messages. These status messages are analysed by JSF, which in turn adjusts the numerical parameters and restarts the calculation. The most frequently encountered numerical problems and their solution are:

Failing convergence of the self-consistent KS cycle is caused by a too large step width (mixing coefficient) during the optimization; lowering the coefficient solves these issues (but slows down the convergence significantly).

Failure when diagonalizing the KS-Hamiltonian in a step of the KS self-consistency cycle can be solved by using the less efficient, but more stable *conjugate gradient* method and/or increasing the size of the basis set and direct-space grids.

Direct space grid dimensions are computed automatically from the cell vectors and the kinetic energy cutoff e_{cut} within QE. However, the resulting grid may not be commensurate with the fractional translation component of non symmorphic symmetry operations, which are in turn marked as non-applicable. While this leads to a mere increase of computation time in the self-consistent calculation, the tetrahedron interpolation scheme as implemented in QE at the time of this writing fails in such cases. JSF overrides QE's automatism in such cases, and explicitely creates direct space grids commensurate with the translation components.

Detection and solution of machine problems

At the beginning of each job, JSF performs tests on the compute nodes assigned to it, including the network interconnection used in parallel calculations, and the storage device used to store large files. Diagnostic messages of the numerical codes are analysed regarding failure of the file system or network interconnection. Furthermore, the actual progress of the numerical codes is monitored in real time: an upper bound for the required time per iteration in the self-consistent KS cycle is established from empirical data (an order of magnitude larger than oberved in any calculation); in case there are any compute nodes that are excessively overloaded or erroneous, this upper bound is violated.

If any of the aforementioned criteria are met, the respective computation job is aborted and rescheduled for future execution; no adjustments are made to the computational parameters or requested resources, as the origin of such errors lies within the machine.

Detection and solution of resource request problems

Messages received from the cluster scheduler are processed in order to detect excessive resource usage, such as computation time or memory. Based on this information, JSF automatically adjusts the resource request and reschedules the job for future execution.

Plausibility and consistence of the generated data

In a final step, the results of each calculation are checked for plausibility and consistence, such as the presence of data for each of the desired energy levels. Moreover, when computing the Fermi bond localization, the distance of the Bader surface to any atom is checked against a threshold, in order to detect the (infrequent) failures in the grid-based Bader volume analysis, while at the same time avoiding spurious contributions from the charge localized at the atomic sites.

A.2.3. Evaluation

The JSF framework, described in this section, was employed to supervise calculations on 8.212 materials within our high-throughput search, and performed very well: as shown in Figure A.2, computational jobs for 2.439 materials *actually were* affected by errors. JSF was in fact able to fully automatically solve the vast majority of the errors: computations

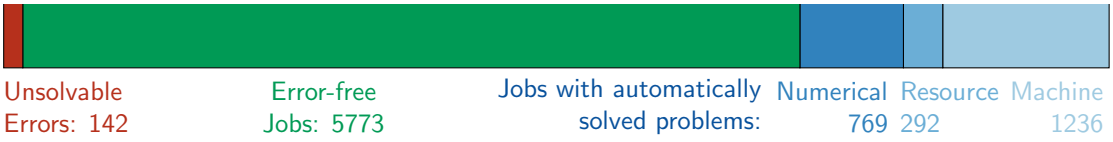


Figure A.2.: Performance of the JSF's automatic error detection and recovery mechanism: each computation job/material is assigned to only one class (sorted by precedence); in case of multiple errors in one job, the error with the highest precedence determines the assignment

succeeded in 2.297 materials due to the error recovery mechanisms. A more in-depth analysis shows that 6.447 errors occured in total, the vast majority (4.965) classified as machine-related, which leads to the conclusion that these 2.297 calculations had a strong overlap with temporary failure of the infrastructure of the computation facility.

A sample of the 142 materials with permanent, i.e. uncorrectable errors was investigated, and showed unrealistic crystal structures, especially in terms of interatomic distances. It is therefore safe to assume that these errors are caused by problematic entries in ICSD.

An analysis of the properties of the 8.069 materials, where calculations were successfull, is presented in chapter 8.

B. Structure maps of superconducting compounds

Within this appendix, the compounds predicted as superconducting by our method are presented (the *elemental solids* were presented separately in subsection 10.4.2, therefore they are not duplicated here).

Basic data on the distribution among binary, ternary and quaternary compounds was outlined in section 10.4, also anticipating a small subset of the information presented within this appendix.

The first section introduces the the concept of representing large numbers of materials via *structure maps*, which are employed for the description of the predicted superconductors in the following sections.

B.1. Repesentation used in the later sections

B.1.1. Subdivision of the set of predicted superconductors

625 compounds in total are predicted to be superconducting by our method. As this number is quite large, too large to be described at once in a way easily accessible to the reader, a split into subsets is performed and each of the subsets will be described separately in the following sections.

As outlined within section 10.4, we choose to subdivide the set of predicted superconducting compounds by

- the number of different elements in their chemical composition (binary, ternary and quaternary systems)
- the stoichiometry of the chemical formula (AB, AB_2, ...)

as such a subdivision allows for a compact representation of the subclasses via *structure maps*, which will be explained in the next subsection.

In case clarity could be improved by it, a further subdivision, is performed based on a chemical classification:

- Intermetallic systems, which do not contain metalloids,
- Intermetallic systems, which do contain metalloids,
- Non-Intermetallic systems.

B.1.2. Structure maps

Structure maps are compact representations of the relation between chemical composition and crystal structure, which were originally introduced by Pettifor [140] and are now commonly used within the high-throughput community.

Each coordinate axis in such a representation is a one-dimensional mapping of the periodic table of elements; two-dimensional coordinates do therefore refer to pairs of chemical elements. Within this work, the *chemical scale* χ from the original reference [140] is used as a 1d-map; χ is a phenomenological scale, which provides an ordering based on approximate chemical similarity. While in principle also the nuclear charge Z does provide a 1d-map of periodic elements, not much insight could be gained from the resulting diagrams, as elements close in Z are not necessarily chemically similar.

Compounds are represented as points in such a coordinate system, where the shape of the point is defined by the *structure prototype* (such as rocksalt or wurzite). In binary compounds, x and y coordinates correspond to the two comprising elements; in ternaries, the third element is represented by the color of a point. Each structure map is accompanied by a classification curve diagram, described in the main body of this thesis (section 10.2), where each subclass material is marked by a point; point shape and color correspond correspond to those used in the structure map. A subset of the classification curve points is explicitely labeled with the chemical formula of the material, highlighting candidates for in-depth studies, either as they are expected to be good superconductors, or are in the direct vincinity of the separator and are good candidates for future refinement of the method. A small subset had been included in the descriptor validation set (section 10.1).

Summarizing, each point in a structure map does fully (modulo pressure-induced variations of the lattice parameters) describe a compound: chemical composition can be read from x, y and optionally color axes, while the shape of each point encodes the crystal structure.

B.2. Predicted superconductors: structure maps

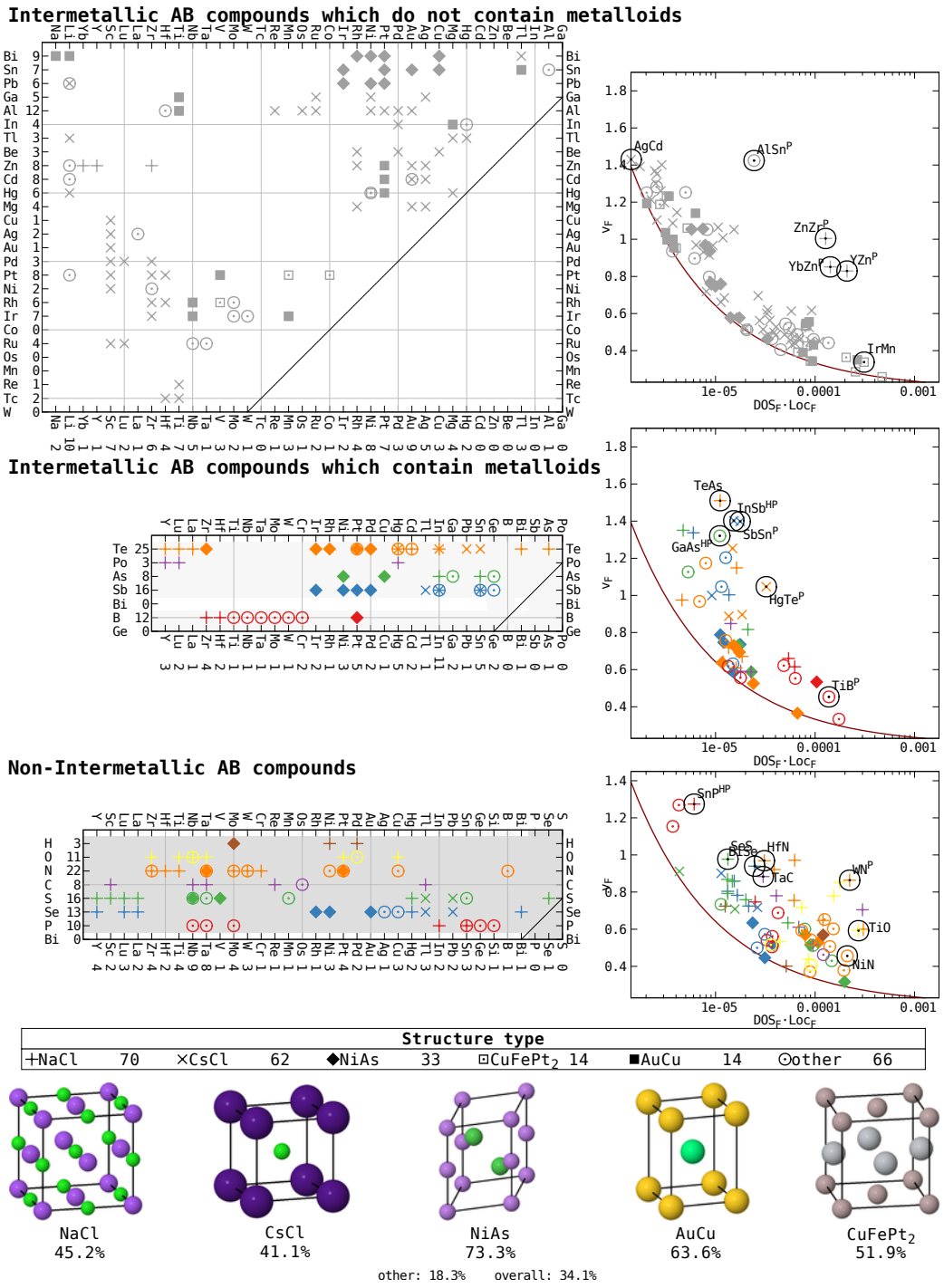


Figure B.1.: Predicted binary superconductors in AB stoichiometry. Compounds are subdivided by chemical composition. Left column: structure map of the subset. Right column: classification diagram Point color for metalloid-intermetallic and not-intermetallic compounds corresponds to element B, i.e. the y coordinate of the structure map. Structure prototypes and point shape legend is presented in lowest part of figure.

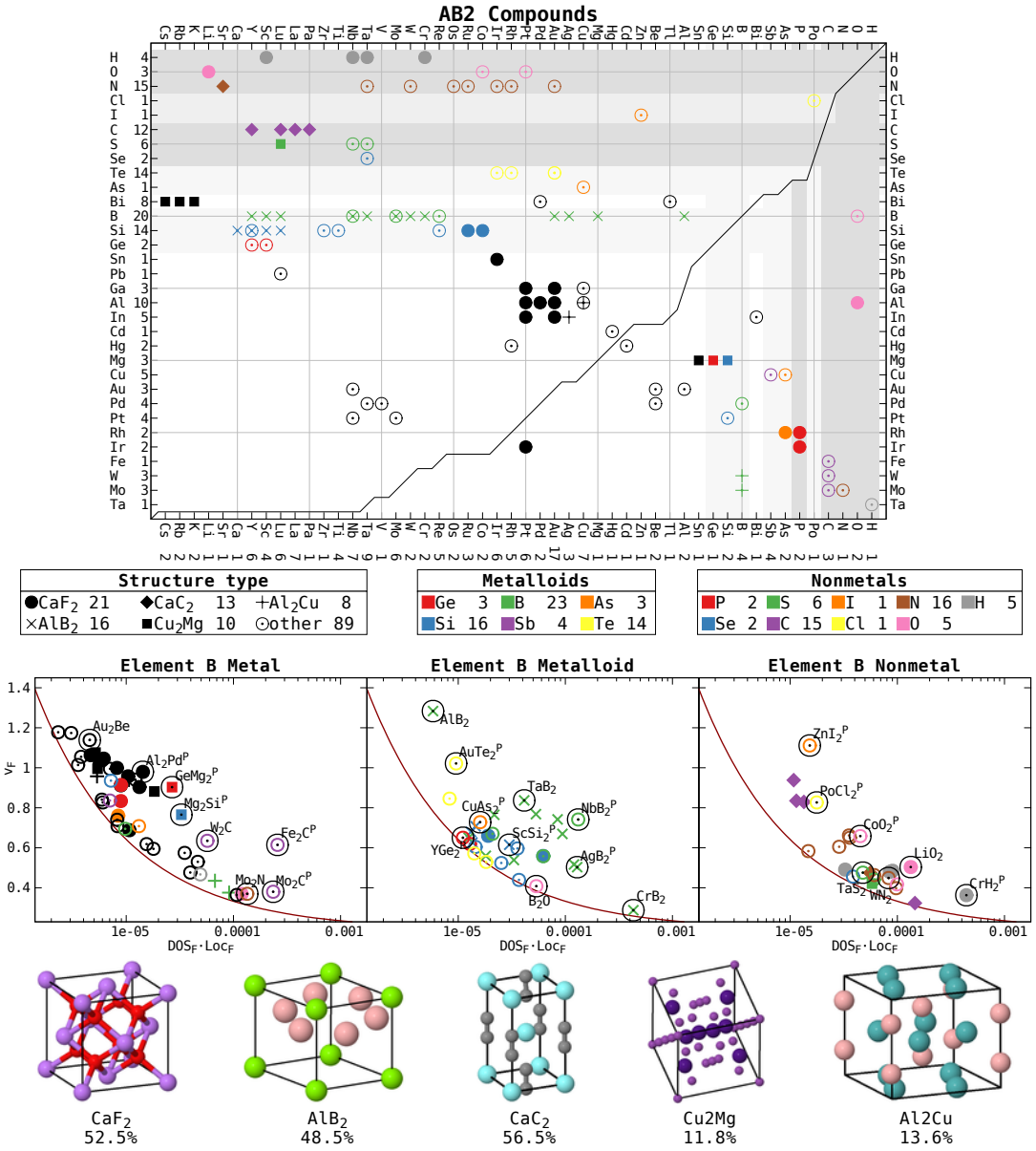


Figure B.2.: Predicted binary superconductors in AB_2 stoichiometry; Top panel: structure map, element A is found on x-axis, element B on y-axis. Point color is either black, or defined by the metalloid/nonmetal atom (legend below structure map). Classification curves are presented in middle row, split by the chemical class of the B element. Final row displays structure prototypes.

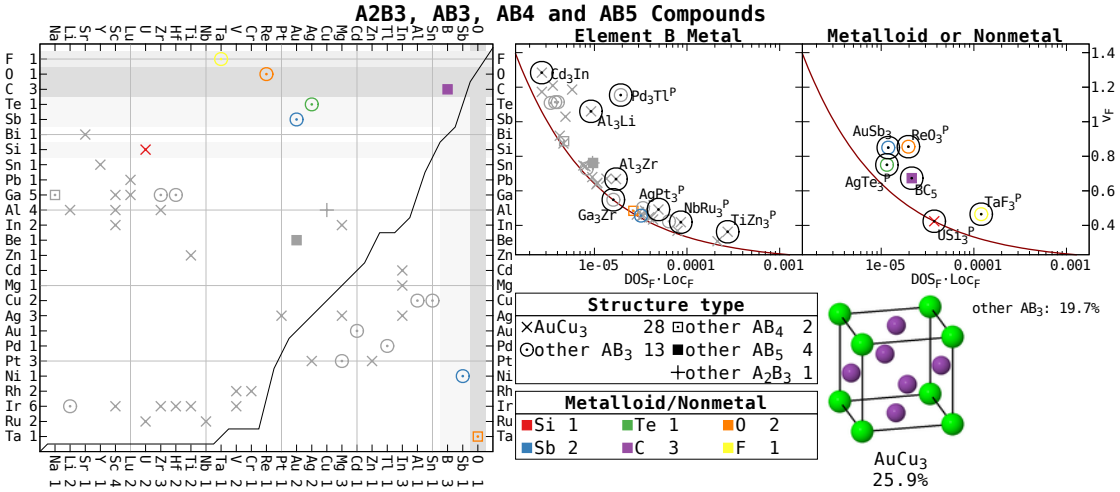


Figure B.3.: Predicted binary superconductors in AB_3, AB_4, AB_5 and A_2B_3 stoichiometries; x coordinate corresponds to element A, y coordinate to element B. Point shape displays stoichiometry and/or structure prototype. Color corresponds to metalloid/nonmetal in chemical composition.

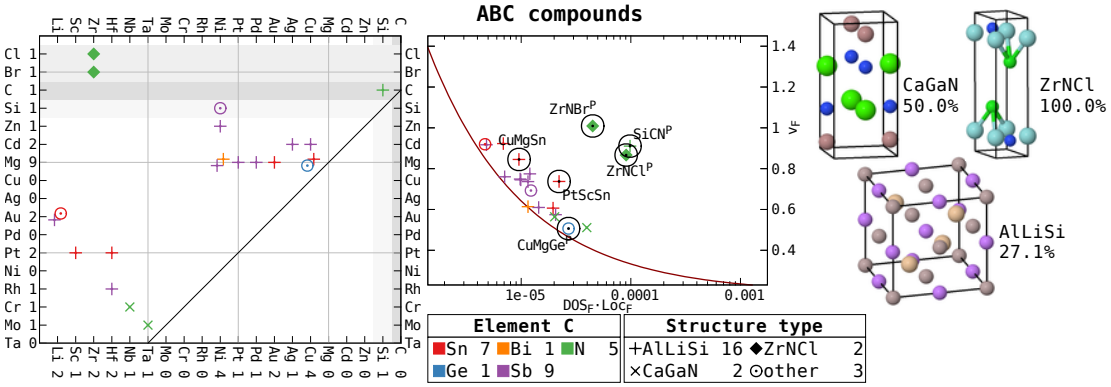


Figure B.4.: Predicted ternary superconductors in ABC stoichiometry. Elements A and B correspond to x and y coordinates, element C determines color.

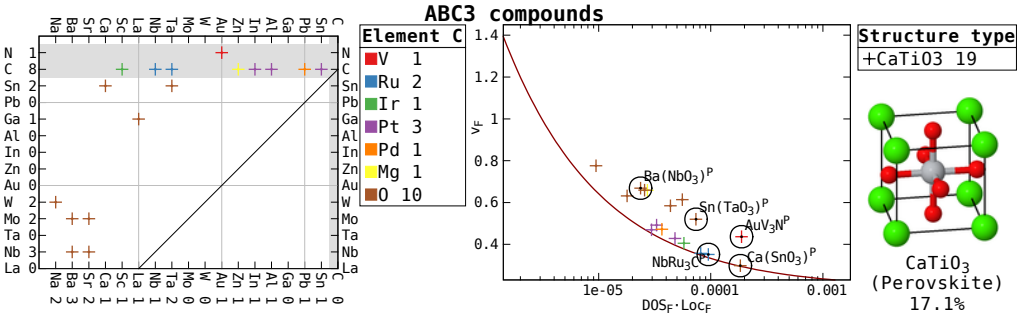


Figure B.5.: Predicted superconductors in ABC_3 stoichiometry. Elements A and B correspond to x and y coordinates, element C determines color.

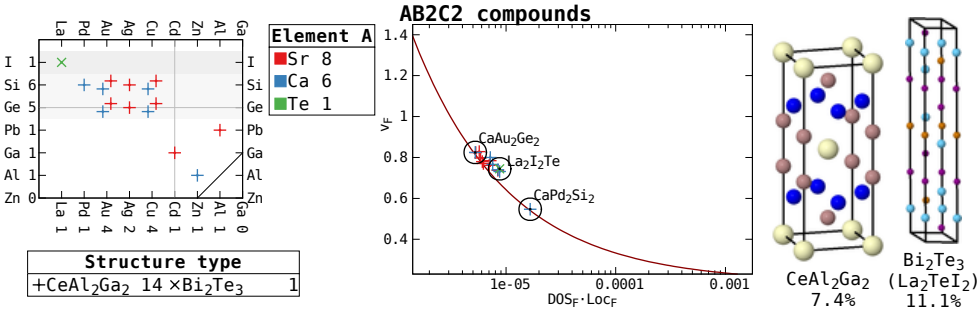


Figure B.6.: Predicted superconductors in AB_2C_2 stoichiometry. Elements B and C correspond to x and y coordinates, element A determines color.

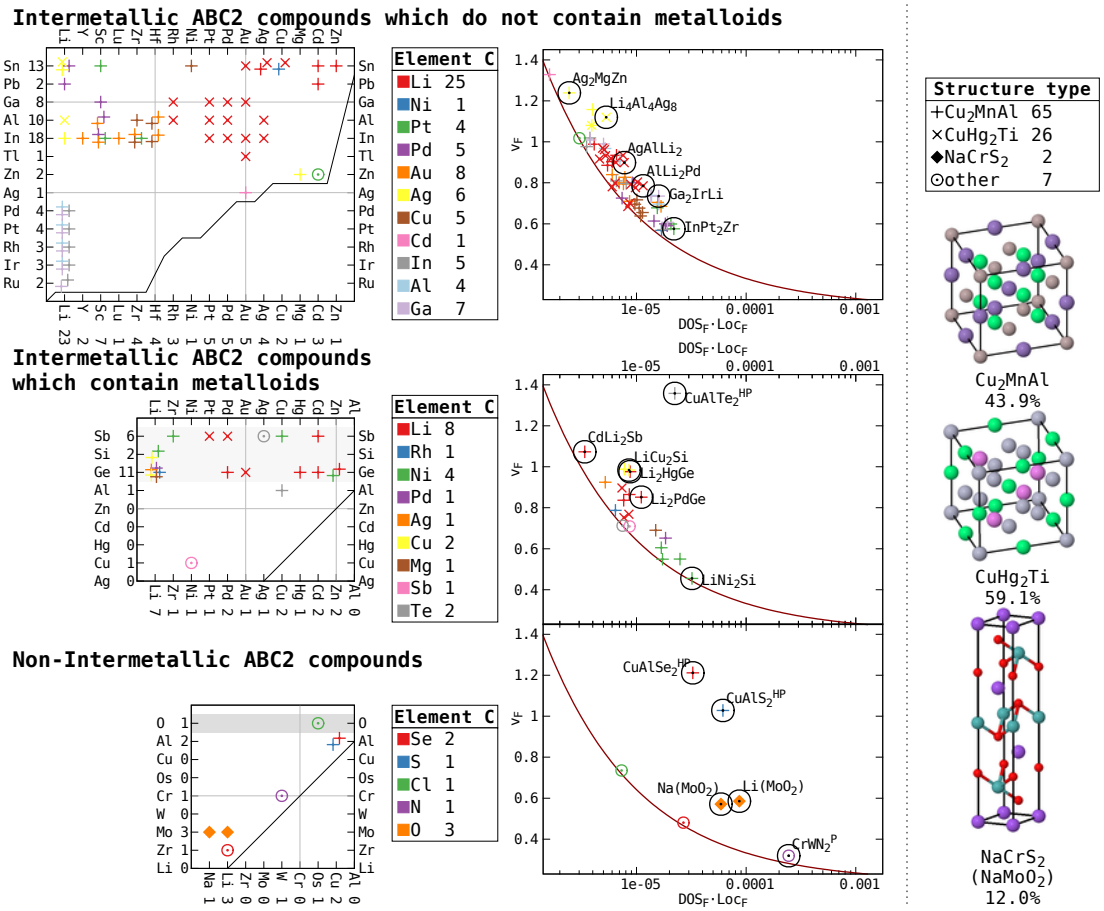


Figure B.7.: Predicted superconductors in ABC_2 stoichiometry. Elements A and B correspond to x and y coordinates, element C determines color.

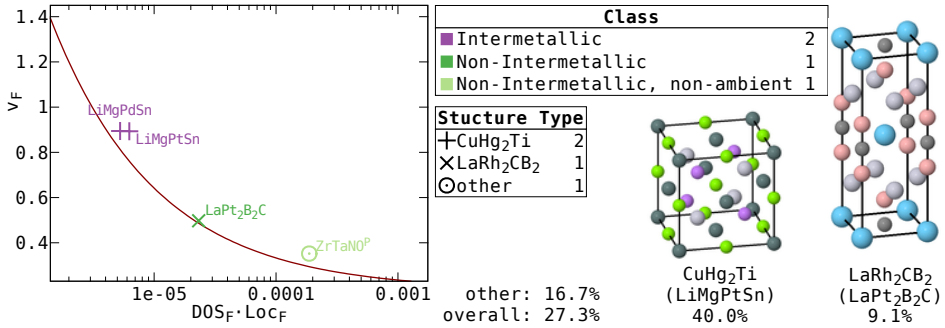


Figure B.8.: Predicted quaternary superconductors. Point shape refers to stoichiometry and/or structure type.

C. Publications

Parts of this thesis have already been published:

1. **How to represent crystal structures for machine learning: Towards fast prediction of electronic properties**, K.T. Schütt*, **H. Glawe***, F. Brockherde, A. Sanna, K.-R. Müller†, E.K.U. Gross†, in *Physical Review B*, **89**: 205118, 2014. DOI: 10.1103/PhysRevB.89.205118

2. **The optimal one dimensional periodic table: a modified Pettifor chemical scale from data mining**, **H. Glawe**, A. Sanna, E.K.U. Gross, M.A.L. Marques, in *New Journal of Physics*, **18**: 093011, 2016. DOI: 10.1088/1367-2630/18/9/093011

*K.T. Schütt and H. Glawe contributed equally to this work.

†Corresponding authors; these authors jointly directed the project.

Bibliography

[1] V. L. Ginzburg. **High-temperature superconductivity-dream or reality?** *Soviet Physics Uspekhi*, **19**: 174, 1976. doi: 10.1070/PU1976v019n02ABEH005136

[2] J. Bednorz and K. Müller. **Possible high Tc superconductivity in the BaLaCuO system**. *Zeitschrift für Physik B Condensed Matter*, **64**: 189–193, 1986. doi: 10.1007/BF01303701

[3] A. Schilling, M. Cantoni, J. D. Guo, and H. R. Ott. **Superconductivity above 130 K in the Hg-Ba-Ca-Cu-O system**. *Nature*, **363**: 56–58, 1993. doi: 10.1038/363056a0

[4] T. Takematsu, R. Hu, T. Takao, Y. Yanagisawa, H. Nakagome, D. Uglietti, T. Kiyoshi, M. Takahashi, and H. Maeda. **Degradation of the performance of a YBCO-coated conductor double pancake coil due to epoxy impregnation**. *Physica C: Superconductivity*, **470**: 674–677, 2010. doi: 10.1016/j.physc.2010.06.009

[5] M. E. Jones and R. E. Marsh. **The Preparation and Structure of Magnesium Boride, MgB2**. *Journal of the American Chemical Society*, **76**: 1434–1436, 1954. doi: 10.1021/ja01634a089

[6] J. Nagamatsu, N. Nakagawa, T. Muranaka, Y. Zenitani, and J. Akimitsu. **Superconductivity at 39K in magnesium diboride**. *Nature*, **410**: 63–64, 2001. doi: 10.1038/35065039

[7] Y. Kamihara, T. Watanabe, M. Hirano, and H. Hosono. **Iron-Based Layered Superconductor La[$O_{1-x}F_x$]FeAs ($x = 0.05 - 0.12$) with $T_c = 26$ K.** *Journal of the American Chemical Society*, **130**: 3296–3297, 2008. doi: 10.1021/ja800073m

[8] J. Yang, Z.-C. Li, W. Lu, W. Yi, X.-L. Shen, Z.-A. Ren, G.-C. Che, X.-L. Dong, L.-L. Sun, F. Zhou, and Z.-X. Zhao. **Superconductivity at 53.5 K in GdFeAs$O_{1-\delta}$**. *Superconductor Science and Technology*, **21**: 082001, 2008. doi: 10.1088/0953-2048/21/8/082001

[9] P. C. Lauterbur. **Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance**. *Nature*, **242**: 190–191, 1973. doi: 10.1038/242190a0

[10] H. K. Onnes. **On the Persistence in Superconductors of Electric Currents without Electromotive Force**. *Scientific American*, **78**: 131–131, 1914.

[11] L. Rossi. **Superconductivity: its role, its success and its setbacks in the Large Hadron Collider of CERN**. *Superconductor Science and Technology*, **23**: 034001, 2010. doi: 10.1088/0953-2048/23/3/034001

[12] N. Mitchell, D. Bessette, R. Gallix, C. Jong, J. Knaster, P. Libeyre, C. Sborchia, and F. Simon. **The ITER magnet system**. *IEEE Transactions on Applied Superconductivity*, **18**: 435–440, 2008. doi: 10.1109/TASC.2008.921232

[13] M. H. Anderson, J. R. Ensher, M. R. Matthews, C. E. Wieman, and E. A. Cornell. **Observation of Bose-Einstein Condensation in a Dilute Atomic Vapor**. *Science*, **269**: 198–201, 1995. doi: 10.1126/science.269.5221.198

[14] C. C. Bradley, C. A. Sackett, and R. G. Hulet. **Bose-Einstein Condensation of Lithium: Observation of Limited Condensate Number**. *Physical Review Letters*, **78**: 985–989, 1997. doi: 10.1103/PhysRevLett.78.985

[15] J. D. Weinstein, R. deCarvalho, T. Guillet, B. Friedrich, and J. M. Doyle. **Magnetic trapping of calcium monohydride molecules at millikelvin temperatures**. *Nature*, **395**: 148–150, 1998. doi: 10.1038/25949

[16] T. Ise, M. Kita, and A. Taguchi. **A hybrid energy storage with a SMES and secondary battery**. *IEEE Transactions on Applied Superconductivity*, **15**: 1915–1918, 2005. doi: 10.1109/TASC.2005.849333

[17] J. P. Stovall, J. A. Demko, P. W. Fisher, M. J. Gouge, J. W. Lue, U. K. Sinha, J. W. Armstrong, R. L. Hughey, D. Lindsay, and J. C. Tolbert. **Installation and operation of the Southwire 30-meter high-temperature superconducting power cable**. *IEEE Transactions on Applied Superconductivity*, **11**: 2467–2472, 2001. doi: 10.1109/77.920363

[18] J. F. Maguire, F. Schmidt, F. Hamber, and T. E. Welsh. **Development and demonstration of a long length HTS cable to operate in the long island power authority transmission grid**. *IEEE Transactions on Applied Superconductivity*, **15**: 1787–1792, 2005. doi: 10.1109/TASC.2005.849289

[19] Y. Li, J. Hao, H. Liu, Y. Li, and Y. Ma. **The metallization and superconductivity of dense hydrogen sulfide**. *The Journal of Chemical Physics*, **140**: 174712, 2014. doi: 10.1063/1.4874158

[20] A. P. Drozdov, M. I. Eremets, and I. A. Troyan. **Conventional superconductivity at 190 K at high pressures**. *ArXiv e-prints*, 2014. arXiv: 1412.0460

[21] H. Rogalla and P. H. Kes. **100 years of superconductivity**. Taylor & Francis, 2011. isbn: 1439849463 doi: 10.1201/b11312

[22] J. P. Devlin. **High throughput screening: the discovery of bioactive substances**. CRC Press, 1997. isbn: 0824700678

[23] M. A. Sills. **Future considerations in HTS: the acute effect of chronic dilemmas**. *Drug Discovery Today*, **3**: 304–312, 1998. doi: 10.1016/S1359-6446(98)01202-1

[24] D. A. Pereira and J. A. Williams. **Origin and evolution of high throughput screening**. *British Journal of Pharmacology*, **152**: 53–61, 2007. doi: 10.1038/sj.bjp.0707373

[25] S. Curtarolo, A. N. Kolmogorov, and F. H. Cocks. **High-throughput ab initio analysis of the BiIn, BiMg, BiSb, InMg, InSb, and MgSb systems**. *Calphad*, **29**: 155–161, 2005. doi: 10.1016/j.calphad.2005.04.003

[26] A. R. Oganov and C. W. Glass. **Crystal structure prediction using ab initio evolutionary techniques: Principles and applications**. *The Journal of Chemical Physics*, **124**: 244704, 244704, 2006. doi: 10.1063/1.2210932

[27] C. J. Pickard and R. J. Needs. **Ab initio random structure searching**. *Journal of Physics: Condensed Matter*, **23**: 053201, 2011. doi: 10.1088/0953-8984/23/5/053201

[28] D. Morgan, G. Ceder, and S. Curtarolo. **High-throughput and data mining with ab initio methods**. *Measurement Science and Technology*, **16**: 296, 2005. doi: 10.1088/0957-0233/16/1/039

[29] K. Kang, Y. S. Meng, J. Bréger, C. P. Grey, and G. Ceder. **Electrodes with High Power and High Capacity for Rechargeable Lithium Batteries**. *Science*, **311**: 977–980, 2006. doi: 10.1002/chin.200620021

[30] H. Chen, G. Hautier, A. Jain, C. Moore, B. Kang, R. Doe, L. Wu, Y. Zhu, Y. Tang, and G. Ceder. **Carbonophosphates: A New Family of Cathode Materials for Li-Ion Batteries Identified Computationally**. *Chemistry of Materials*, **24**: 2009–2016, 2012. doi: 10.1002/chin.201235011

[31] G. Hautier, A. Jain, T. Mueller, C. Moore, S. P. Ong, and G. Ceder. **Designing Multielectron Lithium-Ion Phosphate Cathodes by Mixing Transition Metals**. *Chemistry of Materials*, **25**: 2064–2074, 2013. doi: 10.1021/cm400199j

[32] S. Keinan, M. J. Therien, D. N. Beratan, and W. Yang. **Molecular Design of Porphyrin-Based Nonlinear Optical Materials**. *The Journal of Physical Chemistry A*, **112**: 12203–12207, 2008. doi: 10.1021/jp806351d

[33] R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sanchez-Carrera, L. Vogt, and A. Aspuru-Guzik. **Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics**. *Energy Environ. Sci.*, **4**: 4849–4861, 2011. doi: 10.1039/c1ee02056k

[34] H. Peng, A. Zakutayev, S. Lany, T. R. Paudel, M. d'Avezac, P. F. Ndione, J. D. Perkins, D. S. Ginley, A. R. Nagaraja, N. H. Perry, T. O. Mason, and A. Zunger. **Li-Doped Cr2MnO4: A New p-Type Transparent Conducting Oxide by Computational Materials Design**. *Advanced Functional Materials*, 5267–5276, 2013. doi: 10.1002/adfm.201300807

[35] A. Jain, S.-A. Seyed-Reihani, C. C. Fischer, D. J. Couling, G. Ceder, and W. H. Green. **Ab initio screening of metal sorbents for elemental mercury capture in syngas streams**. *Chemical Engineering Science*, **65**: 3025–3033, 2010. doi: 10.1016/j.ces.2010.01.024

[36] M. Lüders, M. A. L. Marques, N. N. Lathiotakis, A. Floris, G. Profeta, L. Fast, A. Continenza, S. Massidda, and E. K. U. Gross. **Ab initio theory of superconductivity. I. Density functional formalism and approximate functionals**. *Physical Review B*, **72**: 024545, 2005. doi: 10.1103/PhysRevB.72.024545

[37] M. A. L. Marques, M. Lüders, N. N. Lathiotakis, G. Profeta, A. Floris, L. Fast, A. Continenza, E. K. U. Gross, and S. Massidda. **Ab initio theory of superconductivity. II. Application to elemental metals**. *Physical Review B*, **72**: 024546, 2005. doi: 10.1103/PhysRevB.72.024546

[38] A. Floris, A. Sanna, M. Lüders, G. Profeta, N. N. Lathiotakis, M. A. L. Marques, C. Franchini, E. K. U. Gross, A. Continenza, and S. Massidda. **Superconducting properties of MgB_2 from first principles**. *Physica C: Superconductivity*, **456**: 45–53, 2007. doi: 10.1016/j.physc.2007.01.026

[39] C. Bersier, A. Floris, P. Cudazzo, G. Profeta, A. Sanna, F. Bernardini, M. Monni, S. Pittalis, S. Sharma, H. Glawe, A. Continenza, S. Massidda, and E. K. U. Gross. **Multiband superconductivity in Pb, H under pressure and CaBeSi from ab initio calculations**. *Journal of Physics: Condensed Matter*, **21**: 164209, 2009. doi: 10.1088/0953-8984/21/16/164209

[40] C. Bersier, A. Floris, A. Sanna, G. Profeta, A. Continenza, E. K. U. Gross, and S. Massidda. **Electronic, vibrational, and superconducting properties of CaBeSi: First-principles calculations**. *Physical Review B*, **79**: 104503, 2009. doi: 10.1103/PhysRevB.79.104503

[41] S. Massidda, F. Bernardini, C. Bersier, A. Continenza, P. Cudazzo, A. Floris, H. Glawe, M. Monni, S. Pittalis, G. Profeta, A. Sanna, S. Sharma, and E. K. U. Gross. **The role of Coulomb interaction in the superconducting properties of CaC_6 and H under pressure**. *Superconductor Science and Technology*, **22**: 034006, 2009. doi: 10.1088/0953-2048/22/3/034006

[42] G. Profeta, C. Franchini, N. N. Lathiotakis, A. Floris, A. Sanna, M. A. L. Marques, M. Lüders, S. Massidda, E. K. U. Gross, and A. Continenza. **Superconductivity in lithium, potassium, and aluminum**

under extreme pressure: a first-principles study. *Physical Review Letters*, **96**: 047003, 2006. DOI: 10.1103/PhysRevLett.96.047003

[43] A. Sanna, C. Franchini, A. Floris, G. Profeta, N. N. Lathiotakis, M. Lüders, M. A. L. Marques, E. K. U. Gross, A. Continenza, and S. Massidda. **Ab initio prediction of pressure-induced superconductivity in potassium**. *Physical Review B*, **73**: 144512, 2006. DOI: 10.1103/PhysRevB.73.144512

[44] A. Sanna, G. Profeta, S. Massidda, and E. K. U. Gross. **First-principles study of rare-earth-doped superconducting $CaFe_2As_2$**. *Physical Review B*, **86**: 014507, 2012. DOI: 10.1103/PhysRevB.86.014507

[45] J. Bardeen, L. Cooper, and J. Schrieffer. **Microscopic Theory of Superconductivity**. *Physical Review*, **106**: 162–164, 1957. DOI: 10.1103/PhysRev.106.162

[46] J. Bardeen, L. N. Cooper, and J. R. Schrieffer. **Theory of superconductivity**. *Physical Review*, **108**: 1175, 1957. DOI: 10.1063/1.3047438

[47] G. Eliashberg. **Interactions between electrons and lattice vibrations in a superconductor**. *Soviet Physics JETP*, **11**: 696–702, 1960.

[48] K. Miyake, S. Schmitt-Rink, and C. M. Varma. **Spin-fluctuation-mediated even-parity pairing in heavy-fermion superconductors**. *Physical Review B*, **34**: 6554, 1986. DOI: 10.1103/PhysRevB.34.6554

[49] Y. Takada. **Plasmon mechanism of superconductivity in two-and three-dimensional electron systems**. *Journal of the Physical Society of Japan*, **45**: 786–794, 1978. DOI: 10.1143/JPSJ.45.786

[50] A. S. Alexandrov. **Theory of superconductivity: from weak to strong coupling**. CRC Press, 2003. ISBN: 0750308362

[51] **Inorganic Crystal Structure Database (ICSD), Version 2010-2**. Produced by Fachinformationszentrum Karlsruhe Karlsruhe, Germany URL: http://www.fiz-karlsruhe.de/icsd_home.html

[52] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch. **New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design**. *Acta Crystallographica Section B*, **58**: 364–369, 2002. DOI: 10.1107/S0108768102006948

[53] W. Kohn and L. J. Sham. **Self-Consistent Equations Including Exchange and Correlation Effects**. *Physical Review*, **140**: a1133, 1965. DOI: 10.1103/PhysRev.140.A1133

[54] P. Hohenberg and W. Kohn. **Inhomogeneous Electron Gas**. *Physical Review*, **136**: B864, 1964. DOI: 10.1103/PhysRev.136.B864

[55] J. P. Perdew and A. Zunger. **Self-interaction correction to density-functional approximations for many-electron systems**. *Physical Review B*, **23**: 5048, 1981. DOI: 10.1103/PhysRevB.23.5048

[56] H. J. Monkhorst and J. D. Pack. **Special points for Brillouin-zone integrations**. *Physical Review B*, **13**: 5188, 1976. doi: 10.1103/PhysRevB.13.5188

[57] S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi. **Phonons and related crystal properties from density-functional perturbation theory**. *Reviews of Modern Physics*, **73**: 515, 2001. doi: 10.1103/RevModPhys.73.515

[58] E. N. Zein. **Density functional calculations of elastic moduli and phonon spectra of crystals**. *Soviet Physics Solid State*, **26**: 3028–3034, 1984.

[59] S. Baroni, P. Giannozzi, and A. Testa. **Green's-function approach to linear response in solids**. *Physical Review Letters*, **58**: 1861–1864, 1987. doi: 10.1103/PhysRevLett.58.1861

[60] X. Gonze. **Adiabatic density-functional perturbation theory**. *Physical Review A*, **52**: 1096–1114, 1995. doi: 10.1103/PhysRevA.52.1096

[61] E. K. U. Gross, E. Runge, and O. Heinonen. **Many-particle theory**. A. Hilger, 1991. isbn: 0750300728

[62] S. Y. Savrasov, D. Y. Savrasov, and O. K. Andersen. **Linear-response calculations of electron-phonon interactions**. *Physical Review Letters*, **72**: 372–375, 1994. doi: 10.1103/PhysRevLett.72.372

[63] S. Y. Savrasov and D. Y. Savrasov. **Electron-phonon interactions and related physical properties of metals from linear-response theory**. *Physical Review B*, **54**: 16487–16501, 1996. doi: 10.1103/PhysRevB.54.16487

[64] L. Cooper. **Bound Electron Pairs in a Degenerate Fermi Gas**. *Physical Review*, **104**: 1189–1190, 1956. doi: 10.1103/PhysRev.104.1189

[65] H. Fröhlich. **Theory of the Superconducting State. I. The Ground State at the Absolute Zero of Temperature**. *Physical Review*, **79**: 845–856, 1950. doi: 10.1103/PhysRev.79.845

[66] N. N. Bogoliubov. **A new method in the theory of superconductivity. 1.** *Soviet Physics JETP*, **7**: 41–46, 1958. doi: 10.1063/1.3056962

[67] B. D. Josephson. **Possible new effects in superconductive tunnelling**. *Physics Letters*, **1**: 251–253, 1962. doi: 10.1016/0031-9163(62)91369-0

[68] A. B. Migdal. **Interaction between electrons and lattice vibrations in a normal metal**. *Soviet Physics JETP*, **7**: 996–1001, 1958.

[69] Y. Nambu. **Quasi-Particles and Gauge Invariance in the Theory of Superconductivity**. *Physical Review*, **117**: 648–663, 1960. doi: 10.1103/PhysRev.117.648

[70] L. P. Gor'kov. **On the energy spectrum of superconductors**. *Soviet Physics JETP*, **34**: 505–508, 1958.

[71] P. W. Anderson. **Random-Phase Approximation in the Theory of Superconductivity**. *Physical Review*, **112**: 1900–1916, 1958. doi: 10.1103/PhysRev.112.1900

[72] P. Morel and P. W. Anderson. **Calculation of the Superconducting State Parameters with Retarded Electron-Phonon Interaction**. *Physical Review*, **125**: 1263–1271, 1962. doi: 10.1103/PhysRev.125.1263

[73] W. Jones and N. March. **Theoretical Solid State Physics, Vol. 1: Perfect Lattices in Equilibrium**. Dover Books on Advanced Mathematics Dover Publications, 1985. isbn: 0486650154

[74] P. B. Allen and B. Mitrović. **Theory of Superconducting Tc**. in: *Solid State Physics, Vol. 37*. Academic, New York, 1982. 1 doi: 10.1016/S0081-1947(08)60665-7

[75] P. B. Allen and R. C. Dynes. **Transition temperature of strong-coupled superconductors reanalyzed**. *Physical Review B*, **12**: 905, 1975. doi: 10.1103/PhysRevB.12.905

[76] W. L. McMillan. **Transition temperature of strong-coupled superconductors**. *Physical Review*, **167**: 331, 1968. doi: 10.1103/PhysRev.167.331

[77] D. J. Scalapino, J. R. Schrieffer, and J. W. Wilkins. **Strong-coupling superconductivity. I**. *Physical Review*, **148**: 263, 1966. doi: 10.1103/PhysRev.148.263

[78] W. L. McMillan and J. M. Rowell. **Lead phonon spectrum calculated from superconducting density of states**. *Physical Review Letters*, **14**: 108, 1965. doi: 10.1103/PhysRevLett.14.108

[79] R. C. Dynes. **McMillan's equation and the T c of superconductors**. *Solid State Communications*, **10**: 615–618, 1972.

[80] F. S. Khan and P. B. Allen. **Deformation potentials and electron-phonon scattering: Two new theorems**. *Physical Review B*, **29**: 3341, 1984. doi: 10.1103/PhysRevB.29.3341

[81] E. Kartheuser and S. Rodriguez. **Deformation potentials and the electron-phonon interaction in metals**. *Physical Review B*, **33**: 772, 1986. doi: 10.1103/PhysRevB.33.772

[82] K.-R. Müller, S. Mika, G. Ratsch, K. Tsuda, and B. Schölkopf. **An introduction to kernel-based learning algorithms**. *IEEE Transactions on Neural Networks*, **12**: 181–201, 2001. doi: 10.1109/72.914517

[83] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio. **The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects**. *NeuroImage*, **37**: 539–550, 2007. doi: 10.1016/j.neuroimage.2007.01.051

[84] F. Rosenblatt. **The Perceptron, a Perceiving and Recognizing Automaton**. Cornell Aeronautical Laboratory report 1957. 85–460–1

[85] B. E. Boser, I. M. Guyon, and V. N. Vapnik. **A training algorithm for optimal margin classifiers**. in: *Proceedings of the fifth annual workshop on Computational learning theory*. COLT '92 New York, NY, USA: ACM, 1992. 144–152 ISBN: 089791497X DOI: 10.1145/130385.130401

[86] C. Cortes and V. Vapnik. **Support-vector networks**. *Machine Learning*, **20**: 273–297, 1995. DOI: 10.1007/BF00994018

[87] A. Smola and B. Schölkopf. **A tutorial on support vector regression**. *Statistics and Computing*, **14**: 199–222, 2004. DOI: 10.1023/B:STCO.0000035301.49549.88

[88] J. Mercer. **Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations**. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, **209**: 415–446, 1909. DOI: 10.1098/rsta.1909.0016

[89] A. Smola **An introduction to Machine Learning, L3: Perceptron and Kernels**. 2007 URL: http://alex.smola.org/teaching/pune2007/pune_3.pdf

[90] C. Souza **Kernel Functions for Machine Learning Applications**. 2010 URL: http://crsouza.blogspot.de/2010/03/kernel-functions-for-machine-learning.html

[91] C. Saunders, A. Gammerman, and V. Vovk. **Ridge regression learning algorithm in dual variables**. in: *(ICML-1998) Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann 1998. 515–521

[92] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, A. von Lilienfeld, and K.-R. Müller. **Learning Invariant Representations of Molecules for Atomization Energy Prediction**. in: *Advances in Neural Information Processing Systems 25*. 2012. 449–457

[93] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. **Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning**. *Physical Review Letters*, **108**: 058301, 2012. DOI: 10.1103/PhysRevLett.108.058301

[94] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. **Advances in knowledge discovery and data mining**. the MIT Press, 1996. ISBN: 0262560976

[95] M. A. R. Meier, R. Hoogenboom, and U. S. Schubert. **Combinatorial methods, automated synthesis and high-throughput screening in polymer research: The evolution continues**. *Macromolecular Rapid Communications*, **25**: 21–33, 2004. DOI: 10.1002/marc.200300147

[96] A. Jain, G. Hautier, C. J. Moore, S. Ping Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder. **A high-throughput infrastructure for density functional theory calculations**. *Computational Materials Science*, **50**: 2295–2310, 2011. DOI: 10.1016/j.commatsci.2011.02.023

[97] S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehle, H. T. Stokesc, D. O. Demchenkof, and D. Morgang. **AFLOW: an automatic framework for high-throughput materials discovery**. *Computational Materials Science*, **58**: 218–226, 2012. doi: 10.1016/j.commatsci.2012.02.005

[98] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy. **AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations**. *Computational Materials Science*, **58**: 227–235, 2012. doi: 10.1016/j.commatsci.2012.02.002

[99] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson. **The Materials Project: A materials genome approach to accelerating materials innovation**. *APL Materials*, **1**: 011002, 2013. doi: 10.1063/1.4812323

[100] J. Greeley, T. F. Jaramillo, J. Bonde, I. B. Chorkendorff, and J. K. Nørskov. **Computational high-throughput screening of electrocatalytic materials for hydrogen evolution**. *Nature Materials*, **5**: 909–913, 2006. doi: 10.1142/9789814317665_0041

[101] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy. **The high-throughput highway to computational materials design**. *Nature Materials*, **12**: 191–201, 2013. doi: 10.1038/nmat3568

[102] W. E. Pickett. **The other high-temperature superconductors**. *Physica B: Condensed Matter*, **296**: 112–119, 2001. doi: 10.1016/S0921-4526(00)00787-0

[103] I. I. Mazin. **Superconductivity gets an iron boost**. *Nature*, **464**: 183–186, 2010. doi: 10.1038/nature08914

[104] B. T. Matthias. **Transition temperatures of superconductors**. *Physical Review*, **92**: 874, 1953. doi: 10.1103/PhysRev.92.874

[105] B. T. Matthias. **Empirical relation between superconductivity and the number of valence electrons per atom**. *Physical Review*, **97**: 74, 1955. doi: 10.1103/PhysRev.97.74

[106] B. T. Matthias, T. H. Geballe, and V. B. Compton. **Superconductivity**. *Reviews of Modern Physics*, **35**: 1, 1963. doi: 10.1146/annurev.pc.14.100163.001041

[107] B. T. Matthias. **Criteria for superconducting transition temperatures**. *Physica*, **69**: 54–56, 1973. doi: 10.1016/0031-8914(73)90199-7

[108] D. Pines. **Superconductivity in the periodic system**. *Physical Review*, **109**: 280, 1958. doi: 10.1103/PhysRev.109.280

[109] W. E. Pickett and B. M. Klein. **Theory of the normal state heat capacity of Nb3Sn**. in: *Superconductivity in d-and f-band metals 1982*. 1982.

[110] W. E. Pickett. **Transferability and the electron-phonon interaction: A reinterpretation of the rigid-muffin-tin approximation**. *Physical Review B*, **25**: 745–754, 1982. doi: 10.1103/PhysRevB.25.745

[111] S. Saxena, P. Agarwal, K. Ahilan, F. Grosche, R. Haselwimmer, M. Steiner, E. Pugh, I. Walker, S. Julian, P. Monthoux, G. Lonzarich, A. Huxley, I. Sheikin, D. Braithwaite, and J. Flouquet. **Superconductivity on the border of itinerant-electron ferromagnetism in UGe2**. *Nature*, **406**: 587–592, 2000. doi: 10.1016/S0304-8853(00)01324-X

[112] D. Aoki, A. Huxley, E. Ressouche, D. Braithwaite, J. Flouquet, J.-P. Brison, E. Lhotel, and C. Paulsen. **Coexistence of superconductivity and ferromagnetism in URhGe**. *Nature*, **413**: 613–616, 2001. doi: 10.1038/35098048

[113] J. M. An and W. E. Pickett. **Superconductivity of MgB2: covalent bonds driven metallic**. *Physical Review Letters*, **86**: 4366, 2001. doi: 10.1103/PhysRevLett.86.4366

[114] R. F. W. Bader. **Atoms in Molecules: a Quantum Theory**. New York: Oxford University Press, 1990. isbn: 0198558651 doi: 10.1021/ar00109a003

[115] J. Perdew. **Can Density Functional Theory Describe Strongly Correlated Electronic Systems?** English in: *Electron Correlations and Materials Properties 2* ed. by A. Gonis, N. Kioussis, and M. Ciftan. Springer US, 2003. 237–252 isbn: 1441933921 doi: 10.1007/978-1-4757-3760-8_13

[116] M. A. Bösch, M. E. Lines, and M. Labhart. **Magnetoelastic Interactions in Ionic π-Electron Systems: Magnetogyration**. *Physical Review Letters*, **45**: 140–143, 1980. doi: 10.1103/PhysRevLett.45.140

[117] A. K. Nandy, P. Mahadevan, P. Sen, and D. D. Sarma. **KO2: Realization of Orbital Ordering in a p-Orbital System**. *Physical Review Letters*, **105**: 056403, 2010. doi: 10.1103/PhysRevLett.105.056403

[118] I. V. Solovyev. **Spinorbital superexchange physics emerging from interacting oxygen molecules in KO 2**. *New Journal of Physics*, **10**: 013035, 2008. doi: 10.1088/1367-2630/10/1/013035

[119] Z. A. Pardos and N. T. Heffernan. **Using HMMs and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset**. *Journal of Machine Learning Research*, 1, 2010.

[120] Z. Pozoun, K. Hansen, D. Sheppard, M. Rupp, K.-R. Müller, and G. Henkelman. **Optimizing transition states via kernel-based machine learning**. *Journal of Chemical Physics*, **136**: 174101, 2012. doi: 10.1063/1.4707167

[121] J. Behler. **Atom-centered symmetry functions for constructing high-dimensional neural network potentials**. *The Journal of Chemical Physics*, **134**: 074106, 2011. doi: 10.1063/1.3553717

[122] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke. **Finding density functionals with machine learning**. *Physical Review Letters*, **108**: 253002, 2012. doi: 10.1103/PhysRevLett.108.253002

[123] A. P. Bartók, R. Kondor, and G. Csányi. **On representing chemical environments**. *Physical Review B*, **87**: 184115, 2013. doi: 10.1103/PhysRevB.87.184115

[124] O. A. von Lilienfeld, M. Rupp, and A. Knoll. **Representation of molecules as Fourier series of atomic radial distribution functions: A descriptor for machine learning of potential energy surfaces in chemical compound space**. *ArXiv e-prints*, 2013. arXiv: 1307.2918

[125] M. L. Braun, J. M. Buhmann, and K.-R. Müller. **On relevant dimensions in kernel feature spaces**. *The Journal of Machine Learning Research*, **9**: 1875–1908, 2008.

[126] K.-R. Müller **TU Berlin Machine Learning Group**. 2011 url: http://www.ml.tu-berlin.de

[127] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. **Machine learning of molecular electronic properties in chemical compound space**. *New Journal of Physics*, **15**: 095003, 2013. doi: 10.1088/1367-2630/15/9/095003

[128] G. Montavon, M. L. Braun, T. Krueger, and K.-R. Müller. **Analyzing Local Structure in Kernel-Based Learning: Explanation, Complexity, and Reliability Assessment**. *IEEE Signal Processing Magazine*, **30**: 62–74, 2013. doi: 10.1109/MSP.2013.2249294

[129] S. J. L. Billinge and M. F. Thorpe. **Local structure from diffraction**. Springer, 1998. isbn: 0306458276 doi: 10.1007/b119172

[130] G. Forman. **An extensive empirical study of feature selection metrics for text classification**. *The Journal of Machine Learning Research*, **3**: 1289–1305, 2003.

[131] T. Joachims. **Text categorization with Support Vector Machines: Learning with many relevant features**. in: *Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings*. ed. by C. Nédellec and C. Rouveirol Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. 137–142 isbn: 9783540697817 doi: 10.1007/BFb0026683

[132] L. Pauling. **The principles determining the structure of complex ionic crystals**. *Journal of the American Chemical Society*, **51**: 1010–1026, 1929. doi: 10.1021/ja01379a006

[133] L. Pauling. **The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry**. George Fisher Baker Non-Resident Lecture Series Cornell University Press, 1960. 543–562 isbn: 0801403332

[134] S. M. Woodley and R. Catlow. **Crystal structure prediction from first principles**. *Nature Materials*, **7**: 937–946, 2008. doi: 10.1038/nmat2321

[135] D. A. Coley. **An Introduction to Genetic Algorithms for Scientists and Engineers**. Nature Publishing Group, 1999. isbn: 9810236026 doi: 10.1142/3904

[136] S. Hamad, C. R. A. Catlow, S. M. Woodley, S. Lago, and J. A. Mejias. **Structure and stability of small TiO2 nanoparticles**. *Journal of Physical Chemistry B*, **109**: 15741–15748, 2005. doi: 10.1021/jp0521914

[137] D. J. Wales and J. P. K. Doyle. **Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms**. *Journal of Physical Chemistry A*, **101**: 5111–5116, 1997. doi: 10.1021/jp970984n

[138] S. Kirkpatrick, J. C. D. Gellat, and M. P. Vecchi. **Optimization by simulated annealing**. *Science*, **220**: 671–680, 1983. doi: 10.1126/science.220.4598.671

[139] W. Hume-Rothery and G. V. Raynor. **The Structure of Metals and Alloys**. 1956. isbn: 0904357341

[140] D. G. Pettifor. **The structures of binary compounds. I. Phenomenological structure maps**. *Journal of Physics C: Solid State Physics*, **19**: 285, 1986. doi: 10.1088/0022-3719/19/3/002

[141] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder. **Data Mined Ionic Substitutions for the Discovery of New Compounds**. *Inorganic Chemistry*, **50**: PMID: 21142147, 656–663, 2011. doi: 10.1021/ic102031h

[142] L. Yang and G. Ceder. **Data-mined similarity function between material compositions**. *Physical Review B*, **88**: 224107, 2013. doi: 10.1103/PhysRevB.88.224107

[143] T. Moeller. **The Chemistry of the Lanthanides: Pergamon Texts in Inorganic Chemistry**. Elsevier, 2013. isbn: 9781483187631

[144] L. R. Morss, N. M. Edelstein, J. Fuger, and J. J. Katz. **The Chemistry of the Actinide and Transactinide Elements**. vol. 1 Springer Science & Business Media, 2007. isbn: 1402035551

[145] G. Miessler and D. Tarr. **Inorganic Chemistry**. Pearson Education, 2004. isbn: 9780130354716

[146] T. H. Dunning Jr, D. E. Woon, J. Leiding, and L. Chen. **The First Row Anomaly and Recoupled Pair Bonding in the Halides of the Late p-Block Elements**. *Accounts of Chemical Research*, **46**: 359–368, 2012. doi: 10.1021/ar300154a

[147] H. M. Leicester. **Factors Which Led Mendeleev to the Periodic Law**. *Chymia*, 67–74, 1948. doi: 10.2307/27757115

[148] D. Pettifor. **A chemical scale for crystal-structure maps**. *Solid State Communications*, **51**: 31–34, 1984. doi: 10.1016/0038-1098(84)90765-8

[149] G. L. W. Hart, S. Curtarolo, T. B. Massalski, and O. Levy. **Comprehensive Search for New Phases and Compounds in Binary Alloy Systems Based on Platinum-Group Metals, Using a Computational First-Principles Approach**. *Physical Review X*, **3**: 041035, 2013. doi: 10.1103/PhysRevX.3.041035

[150] R. Armiento, B. Kozinsky, M. Fornari, and G. Ceder. **Screening for high-performance piezoelectrics using high-throughput density functional theory**. *Physical Review B*, **84**: 014103, 2011. doi: 10.1103/PhysRevB.84.014103

[151] P. Villars, K. Mathis, and F. Hulliger. **Chapter I - Environment Classification and Structural Stability Maps**. in: *The Structures of Binary Compounds* ed. by J. Hafner, F. Hulliger, W. Jensen, J. Majewski, K. Mathis, P. Villars, and P. Vogl. vol. 2 Cohesion and Structure North-Holland, 1989. 1–103 doi: 10.1016/B978-0-444-87478-8.50005-0

[152] S. Graulis, A. Dakevi, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirs, N. R. Serebryanaya, P. Moeck, R. T. Downs, and A. Le Bail. **Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration**. *Nucleic Acids Research*, **40**: D420–D427, 2012. doi: 10.1093/nar/gkr900

[153] E. Cuthill and J. McKee. **Reducing the Bandwidth of Sparse Symmetric Matrices**. in: *Proceedings of the 1969 24th National Conference*. ACM '69 New York, NY, USA: ACM, 1969. 157–172 doi: 10.1145/800195.805928

[154] P. Pop, O. Matei, and C.-A. Comes. **Reducing the bandwidth of a sparse matrix with a genetic algorithm**. *Optimization*, **63**: 1851–1876, 2014. doi: 10.1080/02331934.2013.830120

[155] E. Scerri. **The Role of Triads in the Evolution of the Periodic Table: Past and Present**. *Journal of Chemical Education*, **85**: 585, 2008. doi: 10.1021/ed085p585

[156] J. P. Perdew. **Density functional theory and the band gap problem**. *International Journal of Quantum Chemistry*, **28**: 497–523, 1985. doi: 10.1002/qua.560280846

[157] T. Hastie, R. Tibshirani, and J. J. H. Friedman. **The elements of statistical learning**. Springer New York, 2001. isbn: 0387848576 doi: 10.1007/978-0-387-84858-7

[158] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Muller. **Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies**. *Journal of Chemical Theory and Computation*, **9**: 3404–3419, 2013. doi: 10.1021/ct400195d

[159] J. A. Flores-Livas and A. Sanna. **Superconductivity in intercalated group-IV honeycomb structures**. *Physical Review B*, **91**: 054508, 2015. doi: 10.1103/PhysRevB.91.054508

[160] S. Kitagawa, H. Kotegawa, H. Tou, H. Ishii, K. Kudo, M. Nohara, and H. Harima. **Pressure-Induced Superconductivity in Mineral Calaverite AuTe2**. *Journal of the Physical Society of Japan*, **82**: 113704, 2013. doi: 10.7566/JPSJ.82.113704

[161] M. Calandra and F. Mauri. **High-Tc Superconductivity in Superhard Diamondlike BC5**. *Physical Review Letters*, **101**: 016401, 2008. doi: 10.1103/PhysRevLett.101.016401

[162] J. E. Moussa, J. Noffsinger, and M. L. Cohen. **Possible thermodynamic stability and superconductivity of antifluorite $Be_2B_xC_{1-x}$**. *Physical Review B*, **78**: 104506, 2008. doi: 10.1103/PhysRevB.78.104506

[163] A. Amalraj, C. Nirmala Louis, and S. R. Gerardin Jayam. **Band Structure, Metallization, And Superconductivity Of GaAs And InAs Under High Pressure**. *Journal of Theoretical and Computational Chemistry*, **06**: 833–843, 2007. doi: 10.1142/S0219633607003416

[164] Y. Timofeev, B. Vinogradov, and E. Yakovlev. **Superconductivity of lead sulfide**. *Soviet Physics Solid State*, **23**: 1474, 1981.

[165] V. B. Begoulev, Y. A. Timofeev, B. V. Vinogradov, and E. N. Yakovlev. **P-T diagrams of lead chalcogenides ($P \leq 35\,\mathrm{GPa}$, $T = 4.2 - 300\,\mathrm{K}$)**. *Fizika Tverdogo Tela*, **31**: 254–256, 1989.

[166] A. A. Raj, C. N. Louis, V. Rejila, and K. Iyakutti. **Band Structure, Metallization And Superconductivity Of InP And InN Under High Pressure**. *Journal of Theoretical and Computational Chemistry*, **11**: 19–33, 2012. doi: 10.1142/S0219633612500022

[167] C. Buzea and K. Robbie. **Assembling the puzzle of superconducting elements: a review**. *Superconductor Science and Technology*, **18**: 1, 2005. doi: 10.1088/0953-2048/18/1/R01

[168] K. Shimizu, T. Kimura, S. Furomoto, K. Takeda, K. Kontani, Y. Onuki, and K. Amaya. **Superconductivity in the non-magnetic state of iron under pressure**. *Nature*, **412**: 316–318, 2001. doi: 10.1038/35085536

[169] C.-J. Kang, K. Kim, and B. I. Min. **Phonon softening and superconductivity triggered by spin-orbit coupling in simple-cubic α-polonium crystals**. *Physical Review B*, **86**: 054115, 2012. doi: 10.1103/PhysRevB.86.054115

[170] F. Essenberger, A. Sanna, A. Linscheid, F. Tandetzky, G. Profeta, P. Cudazzo, and E. K. U. Gross. **Superconducting pairing mediated by spin fluctuations from first principles**. *Physical Review B*, **90**: 214504, 2014. doi: 10.1103/PhysRevB.90.214504

[171] J. Lischner, T. Bazhirov, A. H. MacDonald, M. L. Cohen, and S. G. Louie. **First-principles theory of electron-spin fluctuation coupling and superconducting instabilities in iron selenide**. *Physical Review B*, **91**: 020502, 2015. doi: 10.1103/PhysRevB.91.020502

[172] W. Tang, E. Sanville, and G. Henkelman. **A grid-based Bader analysis algorithm without lattice bias**. *Journal of Physics: Condensed Matter*, **21**: 084204, 2009. doi: 10.1088/0953-8984/21/8/084204

[173] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch. **Quantum Espresso: a modular and open-source software project for quantum simulations of materials**. *Journal of Physics: Condensed Matter*, **21**: 395502, 2009. doi: 10.1088/0953-8984/21/39/395502 url: http://www.quantum-espresso.org

[174] T. Bjorkman **cif2cell**. 2010 url: http://sourceforge.net/projects/cif2cell/

[175] K. Dewhurst, S. Sharma, L. Nordström, F. Cricchio, and F. Bultmark **ELK**. 2014 url: http://elk.sourceforge.net/

[176] H. Glawe. **Density Functional Theory for Superconductors: Application to doped CaCuO2**. Diploma thesis. Freie Universität Berlin, 2006.

[177] **FHI LDA Pseudopotentials**. url: http://www.abinit.org/downloads/psp-links/psp-links/lda_fhi

Zusammenfassung

Diese Arbeit befasst sich mit der Entwicklung theoretischer und numerischer Verfahren, die eine Suche nach neuen Supraleitern per Hochdurchsatz-Methode (HTM) ermöglichen.

HTMs untersuchen tausende Materialien auf eine gewünschte Eigenschaft hin. Die Untersuchung jedes einzelnen Materials erfolgt per numerischer Simulation; daher hat die Rechenzeit jeder einzelnen Simulation einen großen Einfluss auf die Leistungsfähigkeit der HTM. Obwohl es numerische *ab-initio*-Verfahren zur Bestimmung der Supraleitung einzelner Materialien gibt, erfordern diese bei weitem zu viel Rechenzeit, um sie innerhalb einer HTM direkt anzuwenden.

Die Vorhersage neuer Supraleiter gliedert sich in drei Aufgabenstellungen, nämlich die beschleunigte Vorhersage (i) neuer Kristallstrukturen, die thermodynamisch oder zumindest dynamisch stabil sind, und die bisher nicht synthetisiert wurden (ii) einfacher elektronischer Eigenschaften, wie Leitfähigkeit und dem Fehlen magnetischer Instabilität (iii) der Stärke der paarbildenden Wechselwirkung, und damit der Möglichkeit einer hohen kritischen Temperatur T_c; in dieser Arbeit behandeln wir konventionelle Supraleitung, bei der Phononen die vermittelnde Rolle spielen.

Im Kern dieser Arbeit behandeln wir Aufgabenstellung (iii), indem wir *Deskriptoren der Supraleitung* herleiten, die, obwohl ihre Auswertung wenig Zeit erfordert, ausreichend Information bereitstellen, um Materialien für tiefergehende Studien auszuwählen. Unser Ansatz, solche Deskriptoren zu entwickeln, basiert sowohl auf theoretischen Überlegungen als auch auf empirischen Daten, wobei wir den Zusammenhang zwischen beidem anhand von Modellen verdeutlichen. Mit Hilfe unserer numerischen Implementierung führen wir eine Hochdurchsatzsuche durch, die unsere *Deskriptoren* für eine Bibliothek bekannter Materialien auswertet, und identifizieren vielversprechende Kandidaten. Eine Stichprobe untersuchen wir mit rechenzeitintensiven Verfahren, um die *Deskriptoren* zu validieren.

Wir behandeln Aufgabenstellung (ii), indem wir Methoden für maschinelles Lernen (ML) entwickeln, die elektronische Eigenschaften eines Materials direkt anhand der Kristallstruktur vorhersagen. Dazu führen wir eine neue Darstellung von Kristallen ein, die wichtige Symmetrien periodischer Systeme berücksichtigt. Wir werten diese Vorhersagen anhand unserer Hochdurchsatzsuche aus.

Desweiteren behandeln wir Aufgabenstellung (i) und entwickeln eine Methode, um neue Kristallstrukturen durch Substitution chemischer Elemente vorherzusagen. Mit Hilfe einer statistischen Analyse einer grossen Datenbank von Materialien führen wir ein neuartiges Maß der Ähnlichkeit chemischer Elemente ein. Dieses Maß kann bereits als solches genutzt werden, um neue Materialen anhand bekannter vorherzusagen. Wir gehen allerdings noch einen Schritt weiter: basierend auf diesem Maß stellen wir eine neue chemische Skala auf, ähnlich der wohlbekannten Pettifor-Skala, welche eine eindimensionale Ordnung der Elemente anhand ihrer chemischen Ähnlichkeit definiert.

Acknowledgement

During the years it took to develop the idea and complete this thesis, I profited greatly from the support and contributions of many people. It is therefore a pleasant task to express my gratitude to them.

First of all, I would like to thank my supervisor, Prof. E.K.U. Gross, for the opportunity and freedom to follow my research interest. His experience and deep knowledge not only within the field of physics helped me also through hard times and the occasional frustration, which seems to be an integral part of a project such as the present one.

Furthermore I express my gratitude to the members of my PhD committee, especially Prof. F. von Oppen, for taking the time to review this work.

My colleague Antonio Sanna was of great help during this project; his expertise on superconductivity was an inspiration to me, and discussing with him helped me to improve my own knowlege on this challenging topic. I admire his patience when reviewing and proofreading this work. Moreover, I have to thank him and José A. Flores Livas for the contribution of data used in the initial setup of the superconductivity prediction scheme.

My office mates Christophe Bersier and Arkady Davydov were very valueable proofreaders, and discussions with them were a great source of ideas. I am also grateful for the conversations I had with Kay Dewhurst on numerical implementation, which influenced my own one.

Angelica Zacharias supplied a lot of information in the field of chemistry, and without her the respective parts of this work would have suffered.

The part of this work considering machine learning would not have been possible without the fruitful collaboration with Prof. Klaus-Robert Müller, Kristof Schütt and Felix Brockherde, who also introduced me to this interesting topic.

The collaboration with Prof. M.A.L. Marques on the topic of chemical similarity and our revised Pettifor chemical scale gave a new spark to my research, and discussions with him helped in clarifying the presentation of my work.

Moreover, I thank all the members of the Gross research group for the friendly and relaxed environment that provided me with the joy and strength to perform my research and therefore this work.

As I conducted this project at the Max-Planck-Institut für Mikrostrukturphysik in Halle, I would also like to thank all the staff members for their help, especially Ina Goffin, who provided great support in dealing with all the necessary administrative details.